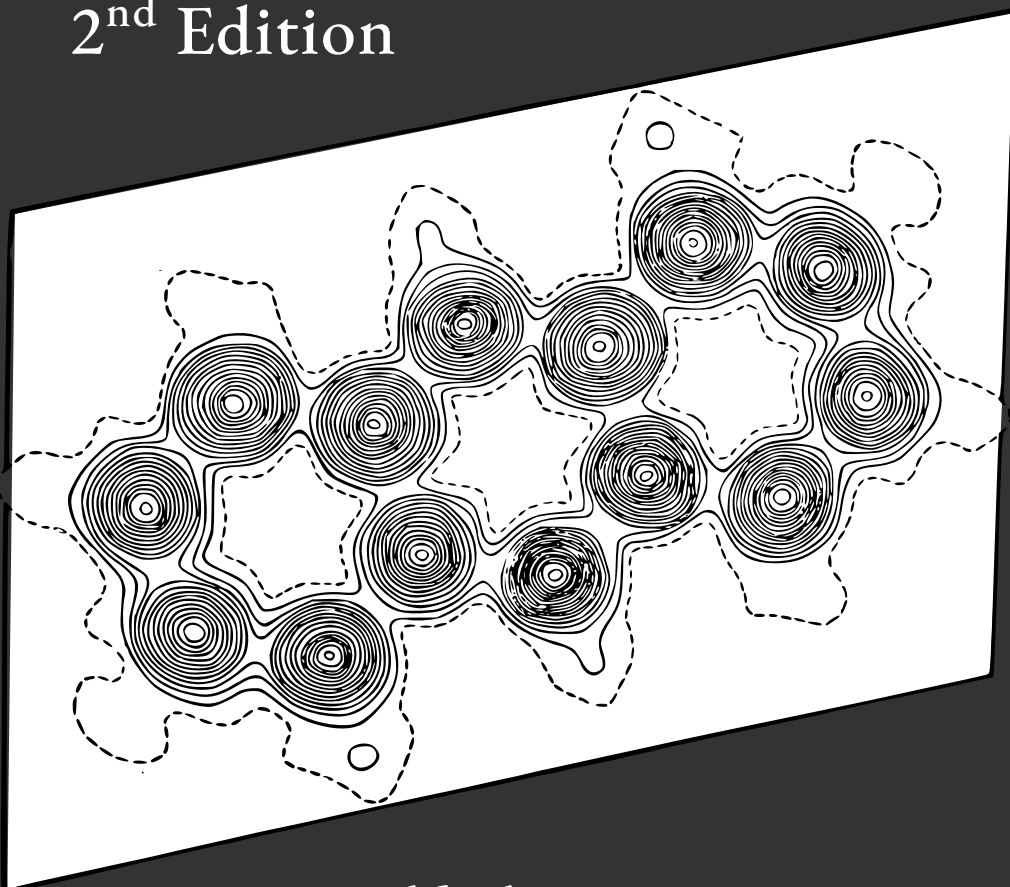


*A. Kitaigorodsky*

# Introduction to Physics

2<sup>nd</sup> Edition



*Mir Publishers Moscow*











А. И. КИТАЙГОРОДСКИЙ

# Введение в физику

ИЗДАТЕЛЬСТВО «НАУКА» МОСКВА

**A. Kitaigorodsky**

# **Introduction to Physics**

Translated from the Russian  
by  
O. Smith and L. Levant

**MIR PUBLISHERS MOSCOW**

First published 1963  
Second printing 1968  
Third printing 1976  
Second edition 1981

*На английском языке*

© Издательство «Наука», Москва

© English translation, Mir Publishers, 1981

# Contents

## Part One

### MECHANICAL AND THERMAL MOTION

<b>Chapter 1. THE FUNDAMENTAL LAW OF MECHANICS</b>	<b>13</b>
Sec. 1. Kinematics . . . . .	13
Sec. 2. Force . . . . .	18
Sec. 3. The Fundamental Law of Mechanics . . . . .	20
Sec. 4. Application of the Fundamental Law of Mechanics to Accelerated Rectilinear Motion . . . . .	22
Sec. 5. Application of the Fundamental Law of Mechanics to Circular Motion . . . . .	25
Sec. 6. The Effect of the Earth's Rotation on Mechanical Phenomena . . . . .	28
Sec. 7. Data Necessary for the Solution of Problems in Mechanics . . . . .	30
Sec. 8. Constants of Proportionality and Dimensions of Physical Quantities . . . . .	32
 <b>Chapter 2. MECHANICAL ENERGY</b>	 <b>34</b>
Sec. 9. Work . . . . .	<b>34</b>
Sec. 10. Kinetic Energy . . . . .	<b>35</b>
Sec. 11. Potential Energy . . . . .	<b>36</b>
Sec. 12. Law of Conservation of Mechanical Energy . . . . .	40
Sec. 13. Potential Curves. Equilibrium . . . . .	42
 <b>Chapter 3. MOMENTUM</b>	 <b>45</b>
Sec. 14. Conservation of Momentum . . . . .	45
Sec. 15. Centre of Mass . . . . .	46
Sec. 16. Collisions . . . . .	47
Sec. 17. Recoil . . . . .	52
 <b>Chapter 4. ROTATION OF A RIGID BODY</b>	 <b>55</b>
Sec. 18. Kinetic Energy of Rotation . . . . .	55
Sec. 19. Moment of Inertia . . . . .	56
Sec. 20. Rotational Work and the Fundamental Equation of Rotation . . . . .	58
Sec. 21. Angular Momentum . . . . .	60
Sec. 22. Free Axes of Rotation . . . . .	62
Sec. 23. The Gyroscope . . . . .	64
 <b>Chapter 5. VIBRATIONS</b>	 <b>65</b>
Sec. 24. Small Deviations from Equilibrium . . . . .	65
Sec. 25. Particular Cases of Vibrations . . . . .	66

Sec. 26. Transformation of Energy. Damped Vibrations . . . . .	68
Sec. 27. Forced Vibrations . . . . .	71
Sec. 28. Self-Sustained Vibrations . . . . .	74
Sec. 29. Addition of Parallel Vibrations . . . . .	76
Sec. 30. Vibration Spectrum . . . . .	78
Sec. 31. Addition of Mutually Perpendicular Vibrations . . . . .	80
 <b>Chapter 6. TRAVELLING WAVES</b>	 <b>82</b>
Sec. 32. Propagation of a Disturbance . . . . .	82
Sec. 33. Generation of Wave Motion . . . . .	84
Sec. 34. Pressure and Velocity of Vibrations . . . . .	86
Sec. 35. Energy Flux . . . . .	87
Sec. 36. Damping of Elastic Waves . . . . .	89
Sec. 37. Interference of Waves . . . . .	90
Sec. 38. Principle of Huygens-Fresnel. Reflection and Refraction of Waves . . . .	92
Sec. 39. Reflection Coefficient . . . . .	94
Sec. 40. The Doppler Effect . . . . .	95
 <b>Chapter 7. STANDING WAVES</b>	 <b>97</b>
Sec. 41. Superposition of Two Waves Travelling in Opposite Directions . . . .	97
Sec. 42. Free Vibrations of a Rod . . . . .	98
Sec. 43. Free Vibrations of Two-Dimensional and Three-Dimensional Systems	100
Sec. 44. Forced Vibrations of Rods and Plates . . . . .	102
Sec. 45. Piezoelectric Vibrations . . . . .	103
 <b>Chapter 8. ACOUSTICS</b>	 <b>105</b>
Sec. 46. The Objective and Subjective Nature of Sound . . . . .	105
Sec. 47. Intensity and Loudness of Sound . . . . .	106
Sec. 48. Architecture and Acoustics . . . . .	108
Sec. 49. The Atmosphere and Acoustics . . . . .	109
Sec. 50. Ultrasonics . . . . .	111
 <b>Chapter 9. TEMPERATURE AND HEAT</b>	 <b>112</b>
Sec. 51. Heat Equilibrium . . . . .	112
Sec. 52. Internal Energy . . . . .	113
Sec. 53. The First Law of Thermodynamics . . . . .	114
Sec. 54. The Internal Energy of Microscopic Systems . . . . .	115
Sec. 55. The Equation of State . . . . .	116
Sec. 56. The Equation of the Gas State . . . . .	118
Sec. 57. The Equations of State of Actual Gases . . . . .	120
 <b>Chapter 10. THERMODYNAMIC PROCESSES</b>	 <b>122</b>
Sec. 58. Graphical Representation . . . . .	122
Sec. 59. Work and Cycles . . . . .	123
Sec. 60. Processes Involving a Change of Gas State . . . . .	124
Sec. 61. The Joule-Thomson Process . . . . .	130
 <b>Chapter 11. ENTROPY</b>	 <b>132</b>
Sec. 62. The Principle of Entropy Existence . . . . .	132
Sec. 63. The Principle of Increasing Entropy . . . . .	134
Sec. 64. The Principle of Operation of a Heat Engine . . . . .	136
Sec. 65. Efficiency of a Carnot Cycle . . . . .	137
Sec. 66. The Second Law of Thermodynamics . . . . .	139



<b>Chapter 12. KINETIC THEORY OF GASES</b>	<b>140</b>
Sec. 67. General	140
Sec. 68. Mean Free Path	141
Sec. 69. Gas Pressure. Root-Mean-Square Velocity of Molecules	142
Sec. 70. Internal Energy of a Gas	145
Sec. 71. Statistical Distribution	146
Sec. 72. Boltzmann's Law	148
Sec. 73. Distribution of Particles with Respect to Height in a Gravitational Field	149
Sec. 74. Velocity Distribution of Molecules	151
Sec. 75. Measurement of the Velocities of Gas Molecules	152
Sec. 76. Probability of a State	153
Sec. 77. Irreversible Processes from the Molecular Viewpoint	155
Sec. 78. Fluctuations. Limits to the Application of the Second Law	156
 <b>Chapter 13. PROCESSES OF TRANSITION TO EQUILIBRIUM</b>	 <b>159</b>
Sec. 79. Diffusion	159
Sec. 80. Thermal Conductivity and Viscosity	160
Sec. 81. Rate of Equalisation	162
Sec. 82. Steady Processes	163
Sec. 83. Motion in a Viscous Medium	165
Sec. 84. Coefficients of Diffusion, Viscosity and Thermal Conductivity for Gases	166
Sec. 85. Ultra-Rarefied Gases	168

## Part Two

### ELECTROMAGNETIC FIELDS

<b>Chapter 14. ELECTRIC FIELDS</b>	<b>170</b>
Sec. 86. Vector Properties of Electric Fields: Intensity and Displacement	170
Sec. 87. Permittivity	171
Sec. 88. Electric Field Relations	173
Sec. 89. Field Calculations of Simple Systems	174
Sec. 90. Electric Energy	183
Sec. 91. Electron Radius and the Limitations of Classical Electrodynamics	185
Sec. 92. Electric Forces	186
Sec. 93. Dipole Moment of a System of Charges	189
Sec. 94. Polarisation of an Isotropic Dielectric	191
Sec. 95. Polarisation of Crystal Substances	193
Sec. 96. Finite Dielectric Bodies in an Electric Field	194
 <b>Chapter 15. MAGNETIC FIELDS</b>	 <b>198</b>
Sec. 97. Magnetic Moment	198
Sec. 98. Ampere Force	200
Sec. 99. Force Acting on a Moving Charge	201
Sec. 100. Magnetic Fields Created by Permanent Magnets	202
Sec. 101. Magnetic Field Intensity	204
Sec. 102. Interactions of Currents and Magnets	206
Sec. 103. Equivalence of Currents and Magnets	207
Sec. 104. Rotational Nature of a Magnetic Field	209
Sec. 105. Law of Electromagnetic Induction and Lorentz Force	212
Sec. 106. Measurement of Magnetic Fields by Means of Induced Impulses	213
Sec. 107. Finite Bodies in a Magnetic Field	215
Sec. 108. Relationship Between Permeability and Susceptibility	218
Sec. 109. Distortion of a Magnetic Field Due to the Presence of a Magnetic Substance	219
Sec. 110. Magnetic Hysteresis	221

<b>Chapter 16. ELECTROMAGNETIC FIELDS. MAXWELL'S EQUATIONS</b>	<b>224</b>
Sec. 111. Generalisation of the Law of Electromagnetic Induction . . . . .	224
Sec. 112. Displacement Current . . . . .	226
Sec. 113. Nature of an Electromagnetic Field . . . . .	229
<b>Chapter 17. ENERGY TRANSFORMATIONS IN ELECTROMAGNETIC FIELDS</b>	<b>231</b>
Sec. 114. Transformations in Steady Current Circuits . . . . .	231
Sec. 115. Transformations in a Closed Circuit of Variable Current . . . . .	232
Sec. 116. Magnetic Energy of a Field . . . . .	234
Sec. 117. Electric Oscillations . . . . .	236
Sec. 118. Electromagnetic Energy . . . . .	238
Sec. 119. Momentum and Pressure of an Electromagnetic Field . . . . .	241
<b>Chapter 18. ELECTROMAGNETIC RADIATION</b>	<b>243</b>
Sec. 120. Elementary Dipole . . . . .	243
Sec. 121. Antennas as Electric Dipoles . . . . .	244
Sec. 122. Radiation Pattern of a Dipole . . . . .	245
Sec. 123. The Electromagnetic Spectrum . . . . .	247
Sec. 124. Quantum Nature of Radiation . . . . .	248
<b>Chapter 19. PROPAGATION OF ELECTROMAGNETIC WAVES</b>	<b>250</b>
Sec. 125. Dispersion and Absorption . . . . .	250
Sec. 126. Behaviour of an Electromagnetic Wave at the Boundary Between Two Media . . . . .	252
Sec. 127. Natural and Polarised Light. Polarisation upon Reflection . . . . .	254
Sec. 128. Propagation of Light Waves in a Medium Having a Refractive Index Gradient . . . . .	255
Sec. 129. Propagation of Radio Waves . . . . .	258
Sec. 130. Radar . . . . .	260
<b>Chapter 20. INTERFERENCE PHENOMENA</b>	<b>262</b>
Sec. 131. Addition of Waves from Two Sources . . . . .	262
Sec. 132. Coherence . . . . .	263
Sec. 133. Interference in a Plate . . . . .	267
Sec. 134. Fringes Representing Equal Thickness and Fringes Representing Equal Inclination . . . . .	268
Sec. 135. Practical Applications of Interference . . . . .	270
<b>Chapter 21. SCATTERING</b>	<b>275</b>
Sec. 136. Secondary Radiation . . . . .	275
Sec. 137. Wave Diffraction at Apertures . . . . .	276
Sec. 138. A System of Randomly Distributed Scatterers . . . . .	279
Sec. 139. Behaviour of a Perfectly Homogeneous Medium . . . . .	281
Sec. 140. Scattering in a Nonhomogeneous Medium . . . . .	282
Sec. 141. Diffraction Grating . . . . .	284
Sec. 142. Directed Radiators of Radio Waves . . . . .	288
Sec. 143. Holography . . . . .	289
<b>Chapter 22. DIFFRACTION OF X-RAYS BY CRYSTALS</b>	<b>292</b>
Sec. 144. Crystals as Diffraction Gratings . . . . .	292
Sec. 145. Determination of the Parameters of a Crystal Cell . . . . .	294

Sec. 146. Intensity of Diffracted Beams . . . . .	295
Sec. 147. Methods of X-Ray Analysis . . . . .	296
<b>Chapter 23. DOUBLE REFRACTION</b>	
Sec. 148. Anisotropic Polarisability . . . . .	299
Sec. 149. Propagation of Light in Uniaxial Crystals . . . . .	301
Sec. 150. Polarisers. Investigation of the Polarised State of Light . . . . .	305
Sec. 151. A Crystal Plate Between "Crossed" Nicol Prisms . . . . .	307
Sec. 152. Double Refraction Due to an External Action . . . . .	308
Sec. 153. Optical Activity . . . . .	310
Sec. 154. Basic Theory of Optical Activity . . . . .	311
<b>Chapter 24. THE THEORY OF RELATIVITY</b>	
Sec. 155. Basic Theory . . . . .	314
Sec. 156. Experimental Verification of the Principle of Constancy of the Velocity of Light . . . . .	315
Sec. 157. Time in the Theory of Relativity . . . . .	318
Sec. 158. Mass . . . . .	319
Sec. 159. Energy . . . . .	320
Sec. 160. Mass Defect . . . . .	321
Sec. 161. The Principle of Equivalence and the General Theory of Relativity . . . . .	321
<b>Chapter 25. THE QUANTUM NATURE OF A FIELD</b>	
Sec. 162. Photons . . . . .	324
Sec. 163. Photoelectric Effect . . . . .	326
Sec. 164. Fluctuations in Luminous Flux . . . . .	327
Sec. 165. Kirchhoff's Law . . . . .	328
Sec. 166. Black-Body Radiation . . . . .	330
Sec. 167. The Theory of Thermal Radiation . . . . .	332
Sec. 168. Stimulated Emission of Radiation . . . . .	335
Sec. 169. Luminescence . . . . .	336

### Part Three

## STRUCTURE AND PROPERTIES OF MATTER

<b>Chapter 26. STREAMS OF CHARGED PARTICLES</b>	
Sec. 170. Motion of Charged Particles in Electric and Magnetic Fields . . . . .	338
Sec. 171. Beams of Charged Particles . . . . .	340
Sec. 172. Electron Lenses . . . . .	341
Sec. 173. The Electron Microscope . . . . .	343
Sec. 174. Electron and Ion Projectors . . . . .	347
Sec. 175. The Electron-Beam Tube . . . . .	348
Sec. 176. Mass Spectrograph . . . . .	350
Sec. 177. Accelerators of Charged Particles . . . . .	352
Sec. 178. Phase Stability . . . . .	353
Sec. 179. Proton and Electron Synchrotrons . . . . .	354
Sec. 180. Ionised Gas . . . . .	355
Sec. 181. Electric Discharges in a Gas . . . . .	356
Sec. 182. Plasma . . . . .	359
<b>Chapter 27. THE WAVE PROPERTIES OF MICROPARTICLES</b>	
Sec. 183. Diffraction of Electrons . . . . .	367
Sec. 184. The Fundamental Concepts of Quantum Mechanics . . . . .	368

Sec. 185. The Uncertainty Principle . . . . .	370
Sec. 186. The Potential Square Well . . . . .	373
Sec. 187. Significance of the Solution of the Schrödinger Equation . . . . .	376
Sec. 188. Tunnelling Through a Barrier . . . . .	377

## Chapter 28. ATOMIC STRUCTURE 379

Sec. 189. Energy Levels of a Hydrogen Atom . . . . .	379
Sec. 190. Quantum Numbers . . . . .	381
Sec. 191. The Electron Cloud of $s$ and $p$ States . . . . .	382
Sec. 192. Pauli's Exclusion Principle . . . . .	383
Sec. 193. Deflection of an Atomic Beam in a Magnetic Field . . . . .	384
Sec. 194. Electron Spin . . . . .	386
Sec. 195. Magnetic Moments of Atoms . . . . .	388
Sec. 196. The Mendeleyev Periodic Law . . . . .	389
Sec. 197. Ionisation Potentials . . . . .	390
Sec. 198. Atomic Spectra in the Optical Region . . . . .	391
Sec. 199. Atomic X-Ray Spectra . . . . .	392

## Chapter 29. MOLECULES 395

Sec. 200. Chemical Bonds . . . . .	395
Sec. 201. Geometries of Molecules . . . . .	397
Sec. 202. The Electronic Cloud of a Molecule . . . . .	399
Sec. 203. Energy Levels of Molecules . . . . .	401
Sec. 204. The Rotational Spectrum of Molecules . . . . .	402
Sec. 205. Infrared Vibration-Rotational Spectra . . . . .	405
Sec. 206. Raman Scattering of Light . . . . .	408
Sec. 207. Absorption Spectra . . . . .	410
Sec. 208. Magnetic Resonance . . . . .	412
Sec. 209. Quadrupole Resonance . . . . .	413
Sec. 210. Gas Lasers . . . . .	415

## Chapter 30. THE ATOMIC NUCLEUS 420

Sec. 211. Experimental Methods of Nuclear Physics . . . . .	420
Sec. 212. Nuclear Particles . . . . .	426
Sec. 213. Mass and Energy of an Atomic Nucleus . . . . .	427
Sec. 214. Spin and Magnetic Moment of a Nucleus . . . . .	429
Sec. 215. Nucleon Interaction Forces . . . . .	430
Sec. 216. Nucleons in a Nucleus . . . . .	432
Sec. 217. Spectra of Atomic Nuclei . . . . .	432
Sec. 218. Neutrino Emitted in Beta-Decay . . . . .	434
Sec. 219. General Laws of Chemical and Nuclear Transformations . . . . .	435
Sec. 220. Radioactivity . . . . .	437
Sec. 221. Nuclear Reactions . . . . .	440
Sec. 222. Fission Reactions of Heavy Nuclei . . . . .	441
Sec. 223. Chain Reactions . . . . .	443
Sec. 224. Nuclear Reactors . . . . .	445
Sec. 225. Artificial Radioactive Products . . . . .	447
Sec. 226. Thermonuclear Reactions . . . . .	448

## Chapter 31. ELEMENTARY PARTICLES 450

Sec. 227. About the Term "Elementary Particle" . . . . .	450
Sec. 228. Interaction of Fast Electrons . . . . .	450
Sec. 229. Meson Theory of Nucleon Interaction . . . . .	451
Sec. 230. Mesons . . . . .	452
Sec. 231. Relativistic Theory of an Electron . . . . .	453
Sec. 232. Creation and Annihilation of Pairs of Particles . . . . .	455
Sec. 233. Particles and Antiparticles . . . . .	456

Sec. 234. Asymmetry of Elementary Particles . . . . .	458
Sec. 235. Baryon Spectrum . . . . .	459
Sec. 236. Quarks . . . . .	462
<b>Chapter 32. ATOMIC STRUCTURE OF BODIES</b>	<b>463</b>
Sec. 237. Polycrystalline Substances and Monocrystals . . . . .	463
Sec. 238. Space Lattice . . . . .	464
Sec. 239. Cell Selection and Crystal Symmetry . . . . .	467
Sec. 240. The Packing of Particles in a Crystal . . . . .	470
Sec. 241. Molecular Crystals . . . . .	471
Sec. 242. Compact Packing of Spheres . . . . .	474
Sec. 243. Examples of Crystal Structures . . . . .	476
Sec. 244. Thermal Vibrations in a Crystal . . . . .	478
Sec. 245. Thermal Waves . . . . .	480
Sec. 246. Thermal Expansion . . . . .	482
Sec. 247. Crystal Imperfections . . . . .	484
Sec. 248. Short-Range Order. Liquids . . . . .	486
Sec. 249. Amorphous Bodies . . . . .	488
Sec. 250. Short- and Long-Range Order of Atoms in Alloys . . . . .	489
Sec. 251. Liquid Crystals . . . . .	491
Sec. 252. Polymers . . . . .	492
Sec. 253. Biological Macromolecules . . . . .	493
<b>Chapter 33. PHASE TRANSFORMATIONS</b>	<b>495</b>
Sec. 254. Phase Diagrams . . . . .	495
Sec. 255. Phase Transformations . . . . .	496
Sec. 256. The Phase Diagram and Properties of Helium . . . . .	497
Sec. 257. Phase Stability . . . . .	500
Sec. 258. Metastable States . . . . .	502
Sec. 259. Gas $\rightleftharpoons$ Liquid Transformations . . . . .	503
Sec. 260. Liquefaction of Gases . . . . .	505
Sec. 261. Gas $\rightleftharpoons$ Crystal Transformations . . . . .	506
Sec. 262. Liquid $\rightleftharpoons$ Crystal Transformations . . . . .	506
Sec. 263. Crystal $\rightleftharpoons$ Crystal Transformations . . . . .	508
Sec. 264. Diffusion in Solids . . . . .	510
<b>Chapter 34. DEFORMATIONS OF BODIES</b>	<b>512</b>
Sec. 265. Elastic Properties . . . . .	512
Sec. 266. Plastic Properties . . . . .	513
Sec. 267. Ultimate Strength . . . . .	515
Sec. 268. Mechanical Properties of a Polycrystalline Material . . . . .	515
Sec. 269. The Effect of Surface-Active Substances on Deformation . . . . .	516
Sec. 270. Material Breakdown Under the Action of a Stream of Particles . . . . .	517
<b>Chapter 35. DIELECTRICS</b>	<b>519</b>
Sec. 271. The Relationship Between Permittivity and the Polarisability of a Molecule . . . . .	519
Sec. 272. Polarisation of Polar and Nonpolar Molecules . . . . .	521
Sec. 273. Additivity of Molecular Refraction . . . . .	524
Sec. 274. Pyroelectric and Piezoelectric Materials . . . . .	525
Sec. 275. Ferroelectric Crystals . . . . .	526
<b>Chapter 36. MAGNETIC SUBSTANCES</b>	<b>529</b>
Sec. 276. Three Groups of Magnetic Substances . . . . .	529
Sec. 277. Diamagnetism . . . . .	529

Sec. 278. Paramagnetism . . . . .	531
Sec. 279. Ferromagnetism . . . . .	532
 Chapter 37. EFFECT OF ELECTRON STRUCTURE ON PROPERTIES OF BODIES	537
Sec. 280. Free Electrons . . . . .	537
Sec. 281. Energy Levels in a Solid . . . . .	538
Sec. 282. Electron Gas . . . . .	540
Sec. 283. Conductivity . . . . .	542
Sec. 284. Superconductivity . . . . .	544
Sec. 285. Semiconductors . . . . .	546
Sec. 286. Emission of Electrons . . . . .	549
Sec. 287. Photoelectric Effect . . . . .	552
Sec. 288. Barrier Layers . . . . .	554
Sec. 289. Contact Potential . . . . .	555
Sec. 290. Electroluminescence of Semiconductors . . . . .	556
Sec. 291. Charge Distribution in a Nonuniformly Heated Body . . . . .	557
Sec. 292. Thermoelectromotive Force . . . . .	558
Sec. 293. Liberation of Heat in Electric Circuits . . . . .	560
Sec. 294. Applications of the Thermoelectric Effect . . . . .	561
Sec. 295. Microelectronic Circuits . . . . .	562
Sec. 296. Technology of Manufacturing Microelectronic Circuits . . . . .	564
Sec. 297. Microprocessors . . . . .	566
Sec. 298. Electronic Computers . . . . .	568
Sec. 299. Electronic Arithmetic . . . . .	569
Sec. 300. Electronic Memory . . . . .	572
 Appendix . . . . .	576
Subject Index . . . . .	581

## PART ONE

# Mechanical and Thermal Motion

### CHAPTER 1

## The Fundamental Law of Mechanics

#### Sec. 1. KINEMATICS

**Equations of Motion of a Particle.** If the dimensions and shape of a body are of no consequence in the consideration of a particular phenomenon, we can conceive of the body as being represented by a point. This approximate representation of a body by a material (i.e., mass) point is not only justified when the dimensions of the body are small relative to other distances considered in the problem, but is permissible whenever we are only interested in the motion of the centre of the mass of the body.

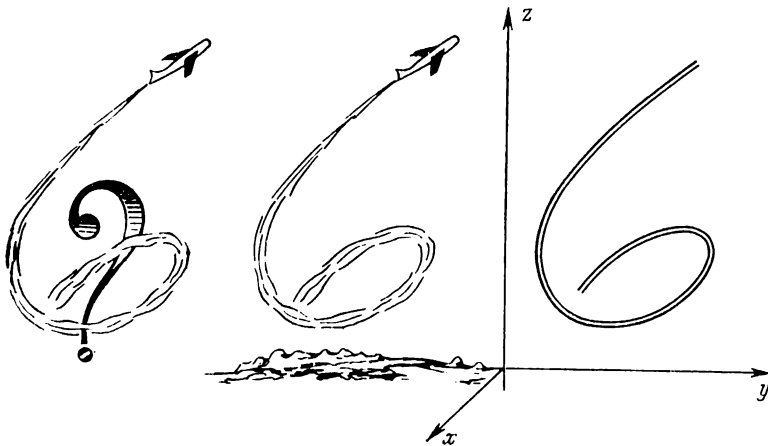


Fig. 1

In order to describe the motion of a particle, one must indicate through which points in space the particle has passed and the instants of time during which it was located at one or another point of the path. For this purpose, it is necessary, in the first place, to select a coordinate frame of reference (Fig. 1). The location of a point in such a coordinate system, which in its simplest form is right-angled, is determined by the three coordinates  $x$ ,  $y$ ,  $z$ , or by the so-called radius vector  $r$ ,

drawn from the origin of the coordinate system to the given point\* (Fig. 2).

Thus, motion in space can be roughly described in the form of a table of values for  $\mathbf{r}$  (each value being given by three quantities!) for the instants of time  $t_1, t_2$ , etc.; or accurately described in the form of a continuous function  $\mathbf{r} = \mathbf{r}(t)$  [in essence, three functions, e.g.,  $x = f_1(t)$ ,  $y = f_2(t)$ ,  $z = f_3(t)$ ; or  $r = \varphi_1(t)$ ,  $\alpha = \varphi_2(t)$ ,  $\beta = \varphi_3(t)$ ; etc.].

The vector equation  $\mathbf{r} = \mathbf{r}(t)$  or, what amounts to the same, the three equivalent scalar equations are called *the equations of motion*.

**Average Velocity.** Let us consider  $AB$ , a portion of the path. Assume that at the instant of time  $t$  the moving particle was at  $A$ , and at the instant of time

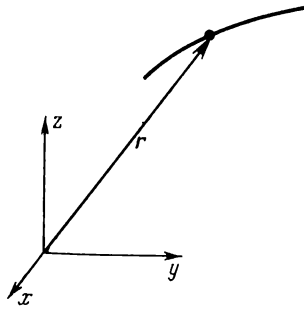


Fig. 2

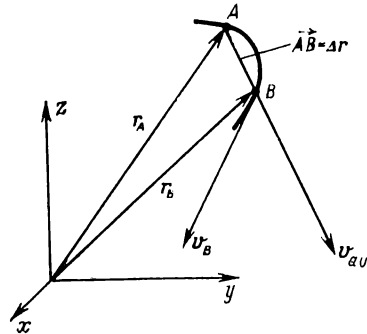


Fig. 3

$t + \Delta t$  at  $B$  (Fig. 3). Let us introduce the radius vectors  $\mathbf{r}_A$  and  $\mathbf{r}_B$ . We know that during the interval of time  $\Delta t$ , the particle moved from  $A$  to  $B$ . It is therefore natural to call the vector  $\vec{AB}$  the particle displacement vector.

Vectors may be added by the parallelogram method. From Fig. 3, we see that

$$\mathbf{r}_B = \mathbf{r}_A + \vec{AB} \quad \text{or} \quad \vec{AB} = \mathbf{r}_B - \mathbf{r}_A = \Delta \mathbf{r},$$

i.e., the particle displacement vector is the vector difference of the radius vectors. The curvilinear motion is determined by the displacement vector  $\Delta \mathbf{r}$  for time  $\Delta t$ , whereby the smaller  $\Delta \mathbf{r}$  the greater the accuracy.

The average speed for the path  $AB$  is given by the relation

$$v_{av} = \frac{AB}{\Delta t}.$$

This is the speed at which the body would have traversed the distance  $AB$  in uniform and rectilinear motion during the interval of time  $\Delta t$ .

Thus, motion over the path  $AB$  may be specified by giving the direction of the vector  $\vec{AB} = \Delta \mathbf{r}$  and the speed  $v_{av}$ . In place of this, we introduce the vector

$$\mathbf{v}_{av} = \frac{\vec{AB}}{\Delta t} = \frac{\Delta \mathbf{r}}{\Delta t},$$

---

\* The radius vector  $\mathbf{r}$  is given by its magnitude,  $r = \sqrt{x^2 + y^2 + z^2}$ , and the angles it forms with the coordinate axes:  $\cos \alpha = \frac{x}{r}$ ,  $\cos \beta = \frac{y}{r}$  and  $\cos \gamma = \frac{z}{r}$ . Thus, it is determined by three quantities:  $x, y$  and  $z$ ; or  $r, \alpha$  and  $\beta$ ; or  $r, \alpha$  and  $\gamma$ ; etc. (two angles determine the third, since  $\cos^2 \alpha + \cos^2 \beta + \cos^2 \gamma = 1$ ).



which is equal in magnitude to the average speed and whose direction is that of the displacement vector. We can now say that the motion of the body over the path  $AB$  is determined by the average velocity.

**Instantaneous Velocity.** If we decrease the interval of time  $\Delta t$ , the point  $B$  will approach point  $A$ . These points finally merge and the direction of  $\vec{AB}$  then coincides with the tangent to the curve at the point of merger.

As  $\Delta t$  decreases, the ratio  $\frac{\vec{AB}}{\Delta t}$  approaches a limit. The vector  $v_{\text{inst}}$ , having the direction of the tangent to the curve at the given moment of motion and numerically equal to the limit of the ratio  $\frac{AB}{\Delta t}$  as  $\Delta t \rightarrow 0$ , is called *the instantaneous particle velocity*:

$$v_{\text{inst}} = \lim_{\Delta t \rightarrow 0} \frac{\Delta r}{\Delta t} \quad \text{when } \Delta t \rightarrow 0.$$

In other words, the **instantaneous velocity** is the derivative of the vector  $r$  with respect to time:

$$v = \frac{dr}{dt}.$$

It should again be emphasised that it is not absolutely essential to employ vectors in order to describe motion. Instead of using the concept of vector velocity, we could speak of the absolute value of the velocity,  $\left| \frac{dr}{dt} \right|$ ,\* and indicate the direction of motion. If we did this, however, the same rules and the same experimental facts would require more cumbersome and more wordy formulations. Vector notation corresponds to physical experience, and is moreover concise and expressive. A certain amount of effort, however, is required to become accustomed to it.

Since the projections of the vector  $r$  on the coordinate axes are the coordinates of its terminus,  $x$ ,  $y$  and  $z$ , the projections of the velocity vector are:

$$v_x = \frac{dx}{dt}, \quad v_y = \frac{dy}{dt} \quad \text{and} \quad v_z = \frac{dz}{dt}$$

**Acceleration.** To continue our consideration of curvilinear motion, let us draw arrows to represent the instantaneous velocities of the body in passing through the points  $A$  and  $B$  of its path. If we had not introduced the concept of velocity, we would have to describe the situation as follows: the speed at  $B$  is different from that at  $A$ ; moreover, the direction of motion has changed. Using the concept of velocity, we can state more briefly: the velocity at  $B$  is different from that at  $A$ .

Velocity can change in magnitude and direction.

If the path  $AB$  is rectilinear, the vectors  $v_A$  and  $v_B$  have the same direction. The change in velocity is obtained by arithmetically subtracting the magnitude of the vector  $v_A$  from the magnitude of the vector  $v_B$ .

Let us now consider the curvilinear path  $AB$ ; vectors  $v_A$  and  $v_B$  differ in magnitude as well as in direction. To determine the increase in the *magnitude* of the velocity, it is necessary, as before, to subtract the magnitude of the vector  $v_A$  from the magnitude of the vector  $v_B$ :

$$\Delta |v| = |v_B| - |v_A|.$$

---

\* The vertical bars  $||$  indicate that only the absolute value (modulus) of the vector between the bars is being considered.

However, this quantity does not, of course, completely express the change that has occurred in the motion.

Let us now subtract vector  $v_A$  from vector  $v_B$  in accordance with the laws of operating on vectors. Fig. 4 shows vector

$$\Delta v = v_B - v_A.$$

Vector  $v_B$ , the sum of  $\Delta v + v_A$ , is the diagonal of the parallelogram constructed on these vectors.

Vector  $\Delta v$  is called the velocity increment. The magnitude of this vector in the case of curvilinear motion is not  $\Delta |v| = |v_B| - |v_A|$ . From the figure it is evident that the magnitude of the increment vector  $|\Delta v|$  is greater than  $\Delta |v|$ , the difference in the magnitudes of the velocities. To determine the velocity at point B, one must add velocity  $v_A$  and increment  $\Delta v$  by the parallelogram method.

We can now determine the acceleration for curvilinear motion as follows. The ratio of the velocity increment to the interval of time during which this increment takes place is called *average acceleration*

$$a_{av} = \frac{\Delta v}{\Delta t}.$$

When the interval of time  $\Delta t$  is decreased, this ratio approaches a limit. The vector

$$a_{inst} = \lim_{\Delta t \rightarrow 0} \frac{\Delta v}{\Delta t} \quad \text{when } \Delta t \rightarrow 0$$

is called the *instantaneous acceleration* of a body at a given moment of motion. In other words, acceleration is the derivative of velocity:

$$a = \frac{dv}{dt}$$

and

$$a_x = \frac{dv_x}{dt}, \quad a_y = \frac{dv_y}{dt}, \quad a_z = \frac{dv_z}{dt}.$$

The acceleration vector uniquely determines the nature of the change in the velocity of the body.

Generally speaking, the acceleration vector can form any angle with the curve. This angle determines the nature of the acceleration and the curvature of the path as follows. Through the point of the curve that is being considered, a circle is drawn that has a common tangent with the path of motion at this point, and for the given portion of the curve most accurately approximates it. This circle is called a *tangential circle*\* and its radius  $\rho$  is called the *radius of curvature* at the given point. The acceleration vector is always directed into this circle. If the motion is accelerated, the vector  $a$  forms an acute angle with the curve (i.e., with the tangent to the path at the given point). If the motion is retarded, this angle will be obtuse. Finally, if the magnitude of the velocity does not change, the acceleration vector is directed normal to the curve.

\* The tangential circle and the calculation of radius of curvature is studied in detail in courses on differential geometry.

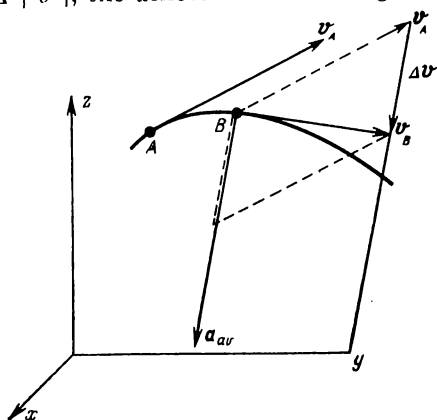


Fig. 4

These statements can be proved rigorously, but we shall merely illustrate them geometrically here (see Fig. 5).

In line with the above discussion, it is customary to resolve the acceleration vector into two components (Fig. 6):

$$\mathbf{a} = \mathbf{a}_t + \mathbf{a}_n.$$

Since the vector triangle is right-angled,

$$a = \sqrt{a_t^2 + a_n^2}.$$

The vector  $\mathbf{a}_t$ , directed along the curve, represents the change in the magnitude of the velocity and is called *the tangential acceleration*. It is not difficult to show that the tangential acceleration

$$a_t = \lim_{\Delta t \rightarrow 0} \frac{\Delta |v|}{\Delta t} \quad \text{when } \Delta t \rightarrow 0,$$

$$\text{i.e., } a_t = \frac{d|v|}{dt},$$

where  $\Delta |v|$  is the increment in the magnitude of the velocity.

The vector  $\mathbf{a}_n$ , directed normal to the curve, represents the change in the direction of the velocity and is called *the normal acceleration*. The normal acceleration  $a_n$  is related by a simple formula to the speed  $v$  and the radius of curvature  $\rho$  at the given point, namely:

$$a_n = \frac{v^2}{\rho}.$$

From this formula, which is derived in courses in theoretical mechanics on the basis of geometrical considerations, it follows that motion with a constant normal acceleration ( $a_n$  and  $v$  constant quantities) is circular motion. In this case,  $\rho$  is a constant quantity for all points along the path and is equal to the radius of the circle.

The normal acceleration  $a_n = \frac{v^2}{\rho}$  is often also called *centripetal acceleration*.

Centripetal acceleration of a body moving in a circle with radius  $R$  can also be expressed by means of the period  $T$ , the frequency  $\nu$ , or the angular velocity  $\omega$  of this motion. Between these quantities and the linear velocity  $v$ , the following simple relations exist:

$$v = \frac{2\pi R}{T}, \quad v = \omega R, \quad v = \frac{1}{T} \quad \text{and} \quad \omega = \frac{2\pi}{T}.$$

The last two formulas define the auxiliary quantities  $\nu$  and  $\omega$ .

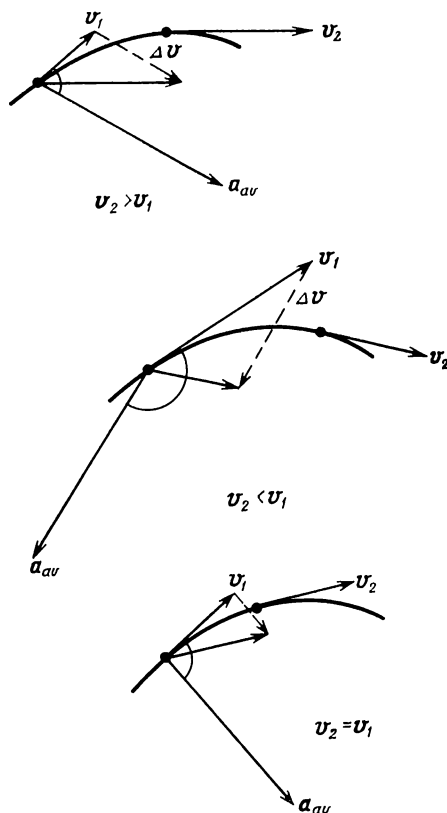


Fig. 5

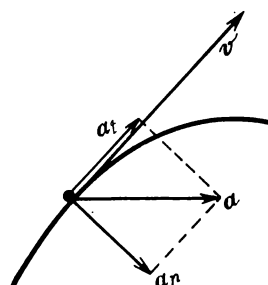


Fig. 6

Thus, the centripetal acceleration for the motion of a body in a circle can also be written in the form:

$$a_n = \omega^2 R \quad \text{or} \quad a_n = \frac{4\pi^2}{T^2} R.$$

It should be emphasised that the everyday understanding of the word "acceleration" is much more limited than in physics. The concept of acceleration in physics includes retardation (negative acceleration) and, what is most important, includes uniform motion if this motion is along a curved path. Only motion that is simultaneously rectilinear and uniform is considered motion without acceleration.

*Examples of acceleration.* A proton in a modern accelerator moves in a circle with a normal acceleration of the order of  $10^{16}$  m/sec<sup>2</sup>. The linear acceleration of a modern rocket is  $\sim 30$  m/sec<sup>2</sup>. The acceleration of a hockey ball is  $\sim 10$  m/sec<sup>2</sup>. The initial acceleration of an automobile is 1-2 m/sec<sup>2</sup>. The angular velocity of the rotor of a turbogenerator is 314 rads/sec and, at a distance of 0.5 metre from the axis of rotation, particles move with an acceleration of  $\sim 5 \times 10^4$  m/sec<sup>2</sup>. The angular velocity of a bicycle wheel is 7-10 rads/sec and, at a radius of 0.5 metre, particles on the rim have a normal acceleration of about 20 m/sec<sup>2</sup>.

## Sec. 2. FORCE

At the present time, four types of interaction are known to the physicist.

**Gravitational Force.** The force of attraction between heavenly bodies, which was discovered by Newton and is otherwise known as gravitational force, acts between any two particles in accordance with the law

$$F = \gamma \frac{m_1 m_2}{r^2},$$

where  $\gamma = 6.67 \times 10^{-11}$  (N·m<sup>2</sup>)/kg<sup>2</sup>,  $m_1$  and  $m_2$  are the masses of the particles, and  $r$  is the distance between them.

It can be rigorously proved, but we shall not do so here, that Newton's law of gravitation written in the form valid for bodies having small dimensions (small with respect to the distance between them) is also valid for the interaction of a small body with a large sphere. The distance, here, is understood to be measured between the centres of the bodies.

The law of universal gravitation for the case of the attraction of a body by the Earth can, therefore, be written in the form

$$F = \gamma \frac{M}{(R+h)^2} m,$$

where  $h$  is the height above the Earth's surface and  $R$  is the radius of the Earth. For points close to the Earth's surface,  $h$  is so much smaller than  $R$  that  $R+h$  may be replaced by  $R$ . Then,  $F = \gamma \frac{M}{R^2} m$ . Comparing this formula with the usual expression for weight,  $F = mg$ , we see that the gravitational acceleration may be expressed in terms of the gravitational constant, the mass of the Earth, and the radius of the Earth:

$$g = \gamma \frac{M}{R^2}$$

Since the gravitational force is proportional to the masses, it is very large for heavenly bodies and negligibly small for the elementary particles. In the interaction between atoms, molecules and other particles of matter, the gravitational force is of no significance.

The force of attraction between the Moon and the Earth is  $2.3 \times 10^{20}$  N, between the Earth and a molecule of oxygen  $\sim 5 \times 10^{-25}$  N, and between two oxygen molecules that are touching each other ( $3 \text{ \AA} = 3 \times 10^{-8}$  cm),  $\sim 2 \times 10^{-42}$  N. These figures speak for themselves.

**Electromagnetic Force.** If two particles or bodies have electric charges  $q_1$  and  $q_2$ , there is a force of attraction between them if the charges are of opposite sign and a force of repulsion if the charges are of equal sign. Quantitatively this relationship is expressed by Coulomb's law:  $F = \frac{q_1 q_2}{r^2}$ . As in the case of universal gravitation, this formula is valid for small particles. We shall show in Sec. 111 that magnetic force and electric force are intimately related. All electromagnetic interaction is of a single nature.

The interaction between atoms, intermolecular forces, and the forces holding electrons about an atomic nucleus are all forces of electrical origin. In order to again demonstrate the negligible nature of the gravitational interaction between elementary particles, we compare gravitational attraction with the electric attraction between a hydrogen atom's nucleus and its single electron:

$$F_{el} = 9 \times 10^{-8} \text{ N, while } F_{grav} = 4 \times 10^{-47} \text{ N!}$$

At first glance, it may not appear understandable why the interaction between neutral atoms and molecules is of electrical origin. We shall go into this in more detail in Chapter 29. However, we should note here that the forces between the atoms and molecules do not depend on the overall charge of the particles (which is equal to zero), but on the local concentration of electric charge.

Since intermolecular force is of electrical origin, surface tension and all cohesion forces between bodies are of the same origin. Frictional force is also essentially based on electric interaction.

The elastic force that is developed when rubber or a compressed metal spring is extended is due to interatomic and intermolecular interaction.

Thus, it too, in the final analysis, is electromagnetic in nature.

**Nuclear Force.** There are forces between neutral particles in an atomic nucleus (also between a neutron and a proton and between two protons) that cannot be explained on the basis of electromagnetism. These forces decrease very rapidly with increasing distance between interacting particles. As a result, these forces do not exist beyond the bounds of nuclei and are evident only in connection with phenomena involving direct interaction of nuclei.

**"Weak" Interaction Force.** These forces appear in the processes of transformation of elementary particles in which neutrinos take part.

**Force Field.** The space in which gravitational force is effective is called a *gravitational field*. Similarly, we speak of an electromagnetic field. Any particle acted on by a force field can also create such a field. Thus, every particle creates a gravitational field and is acted on by gravity; and every electrically charged particle creates an electromagnetic field and is acted on by an electromagnetic field.

Thus, every interaction of particles is depicted in physics according to the scheme: particle—field—particle. The first particle creates a field, and this field acts on the second particle. A few words about how the quantum nature of the field is taken into consideration in this scheme will be said in Sec. 225.

The properties of a field are essentially different from the properties of substance. As a result, it is often said today that matter has two forms—field and substance. The problems of the interrelation of field and substance are at present under intense investigation and cannot, as yet, be considered solved (see p. 186 for a more detailed discussion).

## Sec. 3. THE FUNDAMENTAL LAW OF MECHANICS

**Newton's Laws.** The fundamental law of mechanics is the relationship found by Newton between the forces acting on a body and the acceleration acquired by the body under the action of these forces. This law is usually formulated for particles. This in no way limits the universality of the law, inasmuch as a complex body can be considered, in principle, as the sum total of all its particles. Moreover Newton's equation has extraordinarily broad direct application, since in most problems in mechanics we are either concerned with bodies having small dimensions or are interested only in the motion of the body's centre of mass.

The fundamental law of mechanics states the following. If the forces  $f_1, f_2, f_3$  etc., whose sum total is  $F = \sum f$ , act on a body, the acceleration acquired by the body is equal to the quotient obtained by dividing the resultant force by the mass of the particle:

$$a = \frac{F}{m}.$$

The equation also states that the acceleration vector coincides with the direction of the resultant force. The constant of proportionality in this formula is assumed to be equal to unity, which, the student will recall from his earlier training depends on the choice of the system of units for the quantities entering into this equation.

The fundamental law of mechanics may also be written in the form

$$F = m \frac{dv}{dt}$$

or  $F = \frac{d(mv)}{dt}$ . The latter equation is equivalent to the former only if the mass does not change during the motion. We shall adhere to this condition. The case of variable mass will be considered below. In Chapter 3, we shall briefly discuss the equation of motion for bodies of variable mass in the range typical for rockets; and, in Chapter 24, we shall consider the complications arising when a body moves with a speed approaching that of light (mechanics of the theory of relativity).

The fundamental law of mechanics should be considered as a law generalising observed facts. This equation cannot be theoretically derived from any simple general considerations.

The law of inertia follows directly from the fundamental law. If there are no forces acting on the body, the acceleration is equal to zero and the motion of the body is rectilinear and uniform.

In applying Newton's fundamental law to a particular body, we focus our attention on this body and consider the forces acting on it. It should not be forgotten, however, that force is a measure of the interaction between bodies and that one-sided interaction does not exist. If one body acts on another, the latter also acts on the former. The measurement of force is equivalent to the measurement of interaction. Thus, the very method of measuring force assumes that the force of one body acting on another and the force exerted by the latter on the former are equivalent in magnitude. Since we are usually interested in one particular body, we focus our attention on the force acting on it; the other force is called the force of counteraction or the force of reaction. The forces of action and reaction are equal in magnitude but are oppositely directed. This proposition has become known as Newton's third law of motion.

**Relativity of Motion.** A body at rest in one system of coordinates may appear to us, from another viewpoint, to be moving. The uniform motion of a person

walking along the platform of a station will appear nonuniform if described in a system of coordinates based on a braked train. Therefore, when speaking about the law of motion, the frame of reference for which this law holds must be indicated. The system for which Newton's laws are valid must, without fail, satisfy the following conditions: a body on which no forces are acting must move rectilinearly and uniformly or must be at rest. Such a system is called an *inertial system*.

Thus, it is evident that all frames of reference that are executing accelerated motion **with respect** to a body on which no forces are acting are not inertial systems. Another important conclusion that immediately follows is that there is not merely one inertial system. In fact, an infinite number of inertial systems exist. An inertial system can be based on any body moving uniformly and rectilinearly with respect to some particular body on which no forces are acting.

Let us assume that an inertial system has been selected. Newton's law,  $F = ma$ , is valid for any body moving in this system with a velocity  $v$  and acceleration  $a$ . Now, let us consider another frame of reference moving rectilinearly and uniformly with a velocity  $u$  with respect to the inertial system. To be sure, in this system, the same body will have a different velocity, equal to the difference between the velocity  $v$  and the velocity  $u$ , the motion of the second system with respect to the first. However, since the relative motion of these two systems is rectilinear and uniform, the acceleration of the body will be the same in both systems. Expressed mathematically, since acceleration is the derivative of velocity and the derivative of a constant quantity is equal to zero ( $\frac{du}{dt} = 0$ ):

$$\frac{dv}{dt} = \frac{d(v-u)}{dt}.$$

The acceleration of a body enters into Newton's law, but the velocity does not. As a result, the fundamental law of mechanics is exactly the same in both systems.

This important proposition, following from Newton's law of mechanics, is called *the principle of the relativity of motion*. It can be summarised as follows: An infinite number of inertial systems exist, and, in such systems, the law of inertia and the  $F = ma$  law are satisfied. In this respect, none of these systems has any special advantage over the other systems. All inertial systems are equally suitable for the description of physical phenomena.

The principle of relativity was first formulated by Galileo.

**Laws of Mechanics in a Noninertial System of Coordinates.** Let us assume that the statement "acceleration is due to forces" is always valid in every system of coordinates. In noninertial systems of coordinates, a body executes accelerated motion even when it is not interacting with other bodies. But if this is so, noninertial systems possess, in addition to forces due to interaction, forces of different origin, i.e., forces resulting from the noninertial character of the system. These additional forces are called inertial forces (although it would actually be more correct to call them noninertial forces). Since inertial forces do not result from interaction, they do not satisfy Newton's third law of motion.

We shall limit ourselves to a simple example of inertial force, for in this book we do not intend to employ noninertial systems of coordinates in the analysis of motion.

Let us assume that, for certain reasons, it is convenient to select a system of coordinates moving with an acceleration  $a$ , having a constant magnitude and direction. All bodies at rest or moving uniformly with respect to inertial systems

will move with an acceleration  $-a$  in relation to the noninertial system selected. The acceleration  $-a$  is produced by a force  $-ma$ .

This is the inertial force for the case under consideration. It is not the result of the interaction of bodies, but is due to the accelerated motion of the reference system.

If the body under consideration in the noninertial reference system interacts with other bodies, the inertial force is added to the forces due to interaction.

The fundamental law of mechanics in noninertial systems of coordinates is written in the form:

$$ma = F + \text{inertial forces,}$$

where  $F$  is the resultant force due to the interaction of the bodies.

The expression for the inertial forces will vary in accordance with the nature of the motion of the noninertial reference system (rectilinear, circular, circular with accelerated speed, etc.). Formulas for inertial forces in a variety of cases can be found in books on theoretical physics.

#### Sec. 4. APPLICATION OF THE FUNDAMENTAL LAW OF MECHANICS TO ACCELERATED RECTILINEAR MOTION

In this section, we give several elementary examples illustrating the physical meaning of the fundamental law of mechanics, which states that the vector sum of the forces acting on a body is equal to the product of the mass of the body and the acceleration, and its direction is that of the acceleration.

**Horizontal Motion Under the Action of a Constant Force.** An engine moves a trolley located on rails. Two forces act on the trolley in opposite directions— $F_{rt}$ , the frictional force exerted by the rails, and  $F_{et}$ , the force exerted by the engine. If these two forces are equal, the trolley moves uniformly. In order for the trolley to accelerate, the resultant force must be directed parallel to  $a$ . Therefore, to produce accelerated motion, the motive force must be greater than the frictional force. Moreover, the difference between these forces is the resultant force, which according to the fundamental law of mechanics is equal to the product of the mass and the acceleration. Thus,

$$F_{et} - F_{rt} = ma.$$

The frictional force is the result of the interaction of the rails with the trolley. Therefore, coupled with  $F_{rt}$  is the force exerted on the rails ( $F_{tr}$ ). Similarly, coupled with  $F_{et}$  is  $F_{te}$ , the force which the trolley exerts on the engine.

The force  $F_{te}$  is the resistance force overcome by the engine (experienced by and acting on the latter). This is the force that would act on a man's muscles if he were the source of motive power. As can be seen, the resistance force  $F_{te}$  consists of two terms: the frictional force and the quantity  $-ma$ , which can be called *the inertial resistance*. Inertial resistance is always related to the effective force acting on the accelerated body. It is equal to  $ma$  in magnitude but is oppositely directed to it. The inertial resistance could also be a single force acting on the accelerated body, which would be the case here if there were no friction.

Let us consider another example of horizontal motion under the action of a constant force. The load under consideration is placed on a moving trolley having a back stop (Fig. 7). If there were no back stop, the load could slip off the trolley when the motion is accelerated. With no back stop, the fate of the load depends on the interaction between the trolley's floor and the load. This interaction in-



volves only friction. The trolley moves with a small acceleration  $a$  and the force acting on the load, i.e., the frictional force, should be equal to  $ma^*$ . But the static frictional force cannot have any magnitude whatsoever. It must be somewhat less than  $F_{fr}^{\max}$ . If

$$ma > F_{fr}^{\max},$$

motion with acceleration  $a$  becomes impossible and the load slides off the trolley. If there were no friction between the load and the floor of the trolley, the load would not move from its place, i.e., the trolley would move out from under the load. Let us now assume that the trolley has a back stop. The load is then prevented

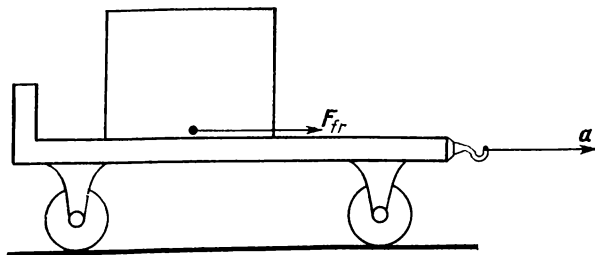


Fig. 7

from sliding as soon as it comes into contact with the back stop. The back stop now pulls the load with the force  $F = ma$ . The force coupled with the motive force is the inertial resistance experienced by the back stop. It is also equal to  $ma$ , but is directed oppositely to the acceleration and acts on the back stop.

*Examples of force.* The force accelerating a passenger car is  $\sim 200$  kgf = 1,960 newtons (N). 1 N is the force imparting an acceleration of 1 metre/sec<sup>2</sup> to a mass of 1 kg; 1 N =  $10^5$  dynes =  $\approx 0.102$  kgf. The thrust developed by the jet engine of a modern aircraft is 10,000–20,000 kgf =  $\approx 10^5$ – $2 \times 10^5$  N and the tractive force developed by a T3-3 diesel locomotive is  $\sim 10,000$  kgf.

**Vertical Motion of an Elevator.** Let us consider the forces acting on a load located on the floor of an elevator executing nonuniform motion.

Assume the elevator is accelerated upward (Fig. 8). Two forces act on the load: the Earth's force,  $F_{El}$ , and the force exerted by the elevator's floor,  $F_{el}$ . But now the resultant force must be different from zero, so  $F_{El} \neq F_{el}$ . Since the resultant force will be in the direction of the acceleration,  $F_{el} > F_{El}$ , and

$$F_{el} - F_{El} = ma.$$

The force  $F_{El}$  is simply the gravitational force exerted by the Earth on the load. Thus,

$$F_{el} - mg = ma.$$

The magnitude of the force exerted by the load on the elevator,  $F_{le}$ , is exactly the same as  $F_{el}$ ; thus, the resistance experienced by the elevator in lifting the load is

$$F_{le} = mg + ma.$$

We see that this resistance consists of the weight of the load and the inertial resistance. The force  $F_{le}$  is sometimes called the apparent weight.

\* In accelerated motion, if some body is carried along merely because of friction (the body carried along is *at rest* with respect to the carrier), the static frictional force will always have the direction of the acceleration.

The result obtained is for the case when the acceleration of the elevator is directed oppositely to that of gravity. This condition is satisfied not only when the elevator is being accelerated upward, but also when it is decelerated during its downward motion.

When the direction of the gravitational force and the acceleration of the elevator coincide, the force exerted by the load on the elevator (apparent weight) is

$$F_{le} = mg - ma.$$

From this formula, it is evident that the pressure against the floor of the elevator ceases when  $a = g$ , i.e., when the elevator falls freely in the gravitational field.

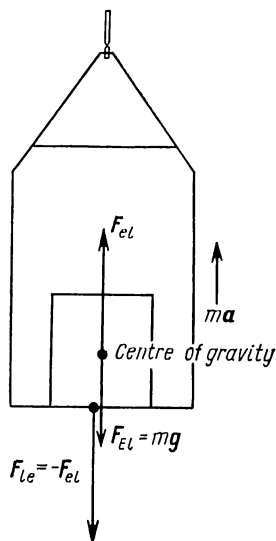


Fig. 8

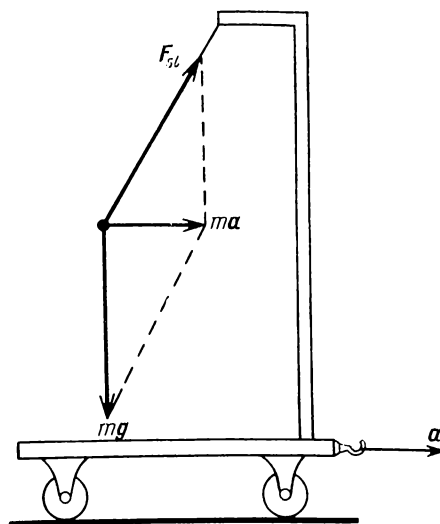


Fig. 9

In this case, the body in the falling elevator ceases to press against the floor and stretch the cable, i.e., the body, so to speak, ceases to have weight.

**Force on a Freely Hanging Load.** Let us consider the motion of a plumb bob suspended from a trolley executing accelerated motion. For such motion, the string by which the plumb bob is suspended forms an angle with the vertical. Two forces act on the load:  $F_{sl}$ , the tension on the string, and  $F_{El}$ , the Earth's attraction, which is equal to  $mg$  (Fig. 9). These forces are directed at an angle to each other. According to the fundamental law of mechanics, their vector sum is equal to  $ma$  and its direction is that of the acceleration. The diagonal of the parallelogram formed by the forces  $F_{sl}$  and  $F_{El}$  is, therefore, horizontal:

$$ma = F_{sl} + F_{El}.$$

The force coupled with  $F_{El}$  is exerted on the Earth and does not interest us. We are, however, interested in the force  $F_{ls}$ , i.e., the force with which the load stretches the string. The value of this force, exerted on the string, is:

$$F_{ls} = -ma + F_{El}.$$

Thus, in this example too, the inertial resistance is a component part of the total resistance experienced by the accelerated body.

### Sec. 5. APPLICATION OF THE FUNDAMENTAL LAW OF MECHANICS TO CIRCULAR MOTION

Motion in a circle is accelerated motion. If a body moves in a circle with constant angular velocity, the magnitude of its acceleration is equal to  $\omega^2 R$  and its direction is inward along the radius.

While executing uniform motion in a circle, the body may be under the action of any number of arbitrarily directed forces. However, in accordance with the fundamental law of mechanics, the vector sum of these forces, or, simply, the resultant force, must be directed inwardly along the radius (parallel to the acceleration) and the value of its magnitude must be

$$F_{\text{centrip}} = \frac{mv^2}{R} = m\omega^2 R.$$

The resultant force acting on the uniformly rotating body is called *the centripetal force*. We again emphasise that the resultant force always has the direction of the acceleration and not of the velocity, i.e., in our case, the force producing uniform circular motion is directed along the radius toward the centre of the circle and not along the tangent to the circular path. The role of the centripetal force is to continually deflect the body from the rectilinear path along which it would move, as the result of inertia, if this force were not present.

*Example.* A particle of mass  $m$  caught on a blade of a modern steam turbine (3,000 rpm, radius about 1 metre) experiences a centripetal force  $F = m\omega^2 r = m \times (314)^2 \times 100 = m \times 10^7$  dynes (where  $m$  is in grams). The weight of the particle is equal to  $mg$ . Thus, the centripetal force is  $\frac{m \times 10^7}{mg}$ , or about 10,000 times, the weight of the particle.

If a body is given accelerated motion, then, in conformity with the law of action and reaction, the accelerated body should act on other bodies (constraints) that make it accelerate rather than move in accordance with the law of inertia. The force of the accelerated body on the constraint has been called the inertial resistance. Such a force also exists, of course for circular motion—it is called *the centrifugal force*.

Centrifugal force and centripetal force are equal in magnitude but are oppositely directed. The centrifugal force is applied to the constraints of a body executing circular motion or, in other words, is applied to those bodies making the body under consideration move in a circle, and preventing it from moving rectilinearly and uniformly. As in the case of centripetal force, centrifugal force is a resultant—the sum of all the reactions exerted by a rotating body on its constraints.

Let us consider several examples, limiting ourselves to the simple case of circular motion due to the interaction of two bodies. If body  $A$  prevents body  $B$  from moving rectilinearly and uniformly, and makes it move uniformly in a circle,  $F_{AB}$  is the centripetal force and  $F_{BA}$  is the centrifugal force. Such simple interaction occurs between a body located on a bowlshaped pedestal, rotating about its axis in the horizontal plane, and the pedestal itself (Fig. 10). If the frictional force is not very large and the pedestal is rotating rapidly, the body slides to the wall of the bowl. In this case, the interaction between the body and the pedestal consists in the following: the wall of the bowl acts on the body inwardly along the radius (centripetal force), and the body, with a force of equal magnitude, presses against the wall outwardly along the radius (centrifugal force).

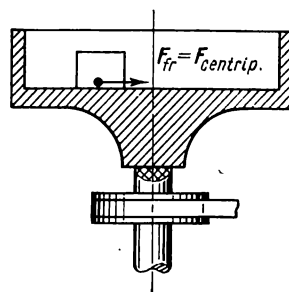


Fig. 10

We return now to the initial moment in this experiment. The body is lying on the pedestal and the pedestal has just begun to rotate. If there were no interaction between the body and the pedestal, the body would remain in place and the pedestal would rotate under the body. The presence of static friction prevents this from happening. The body rotates together with the pedestal. Moreover, as was indicated in the previous section, the static frictional force will be directed inwardly along the radius. The static frictional force is the only force impelling the body to rotate, i.e., the frictional force in this case is a centripetal force. Therefore,

$$F_{fr} = F_{centrip}.$$

The centrifugal force is exerted by the body on the pedestal and is thus directed outwardly along the radius. If for the purpose of clarity (it should be remembered, however, that this is a very gross picture) friction is conceived as being due to the engagement of two rough surfaces, whereby the surface protuberances of the first

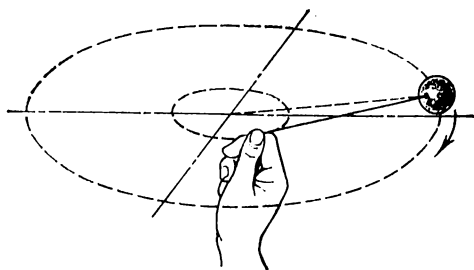


Fig. 11

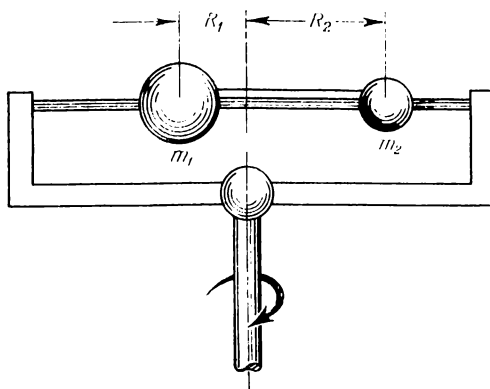


Fig. 12

body mesh with those of the second, then the centrifugal force constitutes a force acting outwardly along the radius at the meshed points of the pedestal surface.

The frictional interaction maintaining the body fixed with respect to the pedestal must be less than a certain maximum  $F_{fr}^{max}$ . In increasing the velocity of rotation of the bowl, the value of  $m\omega^2 R$  finally becomes greater than  $F_{fr}^{max}$ , which makes it impossible for the body to execute circular motion with acceleration  $|a| = \omega^2 R$ . Indeed, to secure circular motion with angular velocity  $\omega$ , a force  $m\omega^2 R$  must act on the body. If the frictional interaction cannot provide this force and, hence, motion in a circle of radius  $R$  with angular velocity  $\omega$ , the body moves with respect to the pedestal and static frictional interaction ceases to exist between the body and the pedestal.

As soon as interaction between the body and the pedestal ceases and the body becomes free, rectilinear and uniform motion begins, the velocity being that possessed by the body at the moment of dissociation. Since the velocity of a body moving in a circle is directed along the tangent, this line represents the line of motion of the freely moving body. The tangential path of particles dissociated from a rotating body is very clearly demonstrated in the case of particles flying from a rotating grindstone.

Let us now consider the rotation of a stone tied to the end of a string (Fig. 11). In order to uniformly rotate the stone at the end of the string under normal condi-

tions, tangential acceleration as well as centripetal acceleration must be imparted to the body. The tangential acceleration is necessary in order to overcome the friction with the air. The resultant acceleration—and, hence, the force—is not directed along the radius, but forms an acute angle with the direction of motion. The hand executes rotational motion and the string is directed at each instant along the tangent to the circle described by the hand.

As another example of circular motion, let us consider the rotation of two attracting bodies having the same angular velocity about a common centre. By means of a centrifugal machine, it is not difficult to make two bodies of equal mass, joined by a string, revolve about a common axis.

To begin with, let us consider the first body, with a string attached to the rotating shaft. The centrifugal force acting on the shaft is equal to  $m_1\omega^2R_1$ . Similarly, the second body acts on the shaft with a force  $m_2\omega^2R_2$ . If these forces are equal, the strings could be joined to each other as shown in Fig. 12, for nothing would change thereby. It is thus clear that the condition for stable rotational motion of two bodies joined by means of a string is the equality of the centrifugal forces exerted on the string by these bodies:

$$m_1\omega^2R_1 = m_2\omega^2R_2.$$

Thus,

$$\frac{m_1}{m_2} = \frac{R_2}{R_1},$$

i.e., stable rotation takes place only when the ratio of the distances to the axis of rotation is inversely proportional to the masses of the bodies.

The point dividing the distance  $R_1 + R_2$  in the ratio  $\frac{R_1}{R_2} = \frac{m_2}{m_1}$  (Fig. 12) is called the centre of mass (see Sec. 15). It can be stated that stable rotation of two joined bodies takes place about the centre of mass of the system.

We have spoken about two bodies whose interaction is achieved by means of a string. However, the above is also completely valid when two bodies attract each other in accordance with the law of universal gravitation, or when a positive and a negative charge attract each other. Thus, interaction of any kind between two attracting bodies can produce stable rotation about the centre of mass of the system. This interaction is given by two forces applied to the attracting bodies. The forces are oppositely directed but numerically equal. (At this point, the unsophisticated reader will usually ask: Why do the bodies not attract each other? We repeat: The forces are parallel to the accelerations, not to the velocities, and in circular motion the acceleration is directed along the radius toward the centre of rotation.) Since a single force acts on each body, both are centripetal forces. At the same time, both are also centrifugal forces. Thus, body *A* acts as a constraint for body *B*, and vice versa. In other words, for body *A*,  $F_{BA}$  is a centripetal force while  $F_{AB}$  is a centrifugal force, and vice versa for body *B*. However, the concept of centrifugal force is used here in a completely formal sense. It was introduced only in order to emphasise the similarity existing between a system of spheres joined by a string and a system of bodies by a force of attraction.

A planetary system is an example of stable rotation of attracting bodies. Let us assume that the Sun had only one planet—the Earth. The centre of rotation would then divide the line joining the Sun and the Earth in the ratio  $m_{\text{Sun}} : m_{\text{Earth}} = 330,000 : 1$ .

Thus, when it is ordinarily said that the Earth rotates about the Sun, we are not committing a serious error, and this would be so even if the Earth were the Sun's only planet.

## Sec. 6. THE EFFECT OF THE EARTH'S ROTATION ON MECHANICAL PHENOMENA

The motion of the Earth is complex. It revolves about its axis and, at the same time, moves in an orbit about the Sun. Hence, it is clear that the Earth does not constitute an inertial frame of reference. Nevertheless, under conditions prevailing on the Earth, Newton's law is generally quite satisfactory. In a number of cases, however, the noninertial property of the Earth's frame of reference has an appreciable effect on the phenomena being studied. These cases should be investigated.

**Effect of the Earth's Rotation on Its Form. Weight of a Body.** If the Earth's rotation is not taken into consideration, a body lying on the surface of the Earth can be considered at rest. The sum of the forces acting on this body would then be equal to zero. As a matter of fact, any particle on the Earth's surface lying at latitude  $\varphi$  moves with an angular velocity  $\omega = 0.7292 \times 10^{-4} \text{ sec}^{-1}$  about the globe's axis, i.e., in a circle of radius  $r = R \cos \varphi$  ( $R$  is the radius of the Earth, which is assumed to have, to a first approximation, the shape of a sphere). There-

fore, the sum of the forces acting on such a particle differs from zero. It is equal to the product of the mass and the acceleration,  $\omega^2 R \cos \varphi$ , and is directed along  $r$ .

It is clear that the presence of such a resultant force  $OG$  (Fig. 13) is possible only when the reaction of the Earth's surface  $OA$  and the gravitational force  $OE$  are directed at an angle to each other. The body will then press on the Earth's surface (according to Newton's third law) with a force  $OC = -OA$ . If the globe were at rest, this force would be equal to the gravitational force  $OE$  and would also coincide with its direction.

Let us resolve the force  $OC$  into two components—one force directed along the radius  $OD$  and the other along the tangent  $OB$ . As can be seen from the figure, the Earth's rotation results in two effects.

First, the weight (the body's pressure on the Earth) becomes less than the gravitational force. Since  $OC \approx OD$ , this decrease equals  $DE = mR\omega^2 \cos^2 \varphi$ . Secondly, a force is produced tending to flatten the Earth, i.e., shift matter toward the equator. This force  $OB = mR\omega^2 \cos \varphi \sin \varphi$ . Such flattening has actually taken place, for the Earth's shape is not spherical but close to an ellipsoid of revolution. As a result of this effect, the equatorial radius of the Earth is  $1/300$  greater than the polar radius.

The flattening force tended to redistribute the mass of the globe as long as the latter's form was not in a state of equilibrium. When this process was completed, the flattening force evidently ceased to be effective. Hence, the force exerted on the terrestrial "globe" is directed normal to the surface.

Let us now return to the quantity expressing the pressure of the body on the Earth, i.e., to the physical quantity generally called weight. The calculation made for a sphere (gravitational force minus  $mR\omega^2 \cos^2 \varphi$ ) is naturally not valid for the actual shape of the Earth. However, for approximate calculations, this value can be used.

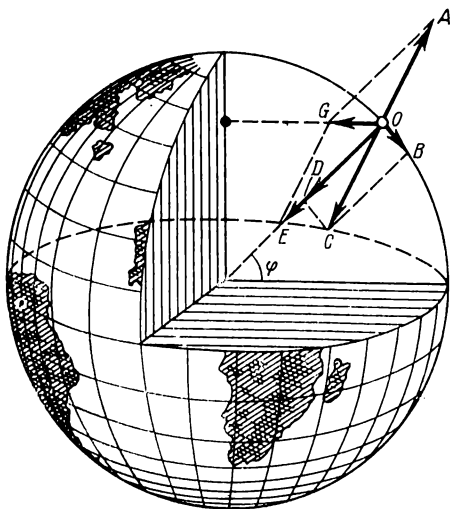


Fig. 13

At the poles ( $\varphi = 90^\circ$ ), the weight of the body is equal to the gravitational force. Let us denote the gravitational force of the body at the poles by  $mg$ . As indicated above, the pressure exerted by the body on the Earth's surface at any point of the globe, in other words, the weight of the body, will be equal to the difference between the gravitational force and the force  $DE$ , i.e.,

$$mg - mR\omega^2 \cos^2 \varphi = mg'.$$

Thus,

$$g' = g - R\omega^2 \cos^2 \varphi$$

is the acceleration with which a body falls at the latitude  $\varphi$ . At the equator,  $g'$  is  $1/300$  less than  $g$ .

If we use an appropriate value for the acceleration of a freely falling body at each of the various latitudes, we do not have to calculate the effect of the Earth's rotation on the weight of the body.

### Effect of the Earth's Rotation on the Motion of a Body on the Earth's Surface.

Let us assume that the motion of a body is observed in a rotating system of coordinates. The body moves rectilinearly and uniformly past the observer, and the motion is curvilinear in the selected noninertial frame of reference. Coriolis, the French scientist, showed by means of calculations that relative to a system rotating with angular velocity  $\omega$  a body moving rectilinearly and uniformly with velocity  $v$  has an acceleration equal to  $2v\omega \sin \alpha$ , where  $\alpha$  is the angle between the axis of rotation and the direction of the rectilinear motion. The acceleration is directed perpendicular to the plane passing through the

axis of rotation and the direction of the velocity. We may use the following rule to determine which of the two possible directions is that of the acceleration. If one looks along the axis of rotation in the direction that makes the rotation appear counterclockwise and places his left hand palm down with the fingers pointing in the direction of the rectilinear motion, the thumb will point in the direction of the acceleration (Fig. 14).

The Coriolis acceleration  $a_{cor}$  acts on all bodies moving on the Earth's surface. If one looks along the Earth's axis from the North Pole, the rotation appears counterclockwise. Hence, in the Northern Hemisphere, any body moving rectilinearly relative to an inertial system will deviate to the right (as viewed by a terrestrial observer) in the course of its motion, while in the Southern Hemisphere, it will deviate to the left. This deviation could be large or small, depending on the direction of motion with respect to the axis and on the linear velocity of the motion.

The deviation of the body can take place in the horizontal or in the vertical plane (with respect to the surface of the Earth). The Coriolis acceleration is directed perpendicular to the Earth's axis; hence, the deviation taking place in the horizontal plane is greatest at the poles and equal to zero at the equator. The reverse is true for deviations in the vertical plane. The deviations in these two planes determine the corresponding projections of the acceleration vector. Thus, the projection of the acceleration of the body in the horizontal plane is

$$2v\omega \sin \varphi,$$

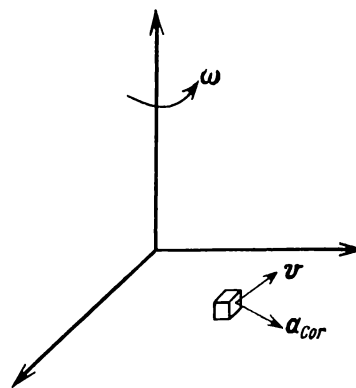


Fig. 14

where  $\varphi$  is the latitude. In the Northern Hemisphere, this projection is directed to the right of the motion.

The deviation of bodies moving in the horizontal plane from their rectilinear path is the reason why the right banks of rivers are eroded in the Northern Hemisphere, and the left banks in the Southern Hemisphere. For the same reason, rivers in the Northern Hemisphere by-pass obstacles to the right and in the Southern Hemisphere to the left.

Air masses flowing into regions of low pressure deviate from the radial direction to the right in the Northern Hemisphere (to the left, in the Southern Hemisphere) and form cyclones. Thus, cyclones in the Northern Hemisphere stir the air masses counterclockwise, and in the Southern Hemisphere clockwise.

As a result of the vertical deviation, a falling body does not fall exactly vertically. Such a body deviates from east to west (the Earth rotates from west to east, i.e., counterclockwise if viewed from the North Pole).

*Examples.* 1. Let us calculate the maximum deviation of a normal artillery shell. The deviation will be a maximum at the poles ( $\varphi = 90^\circ$  and for all firing directions  $\alpha = 90^\circ$ ). If we take the velocity of the shell to be 1 km/sec, we obtain a deviation of  $2 \times 1,000 \times 0.73 \times 10^{-4} \approx 0.15$  m/sec<sup>2</sup>. The gravitational acceleration is about 70 times greater than this acceleration. As can be seen,<sup>†</sup> the deviation of the shell from its rectilinear path can attain a magnitude of the order of several centimetres per second.

2. Assume that a river is flowing from north to south (in the Northern Hemisphere) with a velocity  $v = 3$  km/hr. Thus, the water moves from a region of small linear velocity of rotation of the Earth's surface to a region of larger linear velocity. This increase in the velocity of motion (directed from west to east, together with the banks of the river) is determined by the Coriolis acceleration and is due to the action of the right bank of the river on the mass of water. Let us calculate the Coriolis acceleration for the latitude  $\varphi = 45^\circ$ :

$$a_{Cor} = 2v\omega_0 \sin \varphi,$$

$$\omega_0 = 2\pi \text{rads/day} = 7.25 \times 10^{-5} \text{ rad/sec}, \quad v = 3 \text{ km/hr} = 0.83 \text{ m/sec},$$

$$a_{Cor} = 2 \times 0.83 \times 7.25 \times 10^{-5} \times 0.707 = 8.50 \times 10^{-5} \text{ m/sec}^2.$$

Thus, on every ton of water the right bank exerts a force of

$$8.5 \times 10^{-5} \times 10^3 = 8.5 \times 10^{-2} \text{ N}.$$

The steep, right banks of the Volga, Don and other big rivers of the Northern Hemisphere illustrate this effect.

## Sec. 7. DATA NECESSARY FOR THE SOLUTION OF PROBLEMS IN MECHANICS

The basic problem of mechanics is the determination of the motion for given forces. To determine the motion means to be able to indicate the location in space and the corresponding instant of time for any of the particles. If we are concerned with a complex mechanical system, then such data are necessary for each of the particles into which this system can be considered divided.

In order to tackle such a problem, we must, in the first place, have complete data on the effective forces. The forces must be known for every particle and for every location of this particle. If these forces are known, then by means of Newton's equations we can determine the acceleration of the particle. However, Newton's equations of motion alone are insufficient to completely determine the path, the velocity and the instant of time corresponding to the passage through a given point in space. To describe the motion, it is necessary to know for each instant of time the location of the particle and the magnitude and direction of the velocity. In all, six quantities must be given: three coordinates and the three projections of the velocity on the axes. These data uniquely describe the "mechanical state" of a particle and may be called the parameters of state.



Thus, the problem reduces to the determination of the parameters of state, for Newton's equations only give the acceleration.

To solve the problem, the initial conditions must be known, i.e., the values of the parameters of state for some instant of time (this instant is usually designated by  $t = 0$ , whence the designation "initial conditions"). If the initial values of the parameters of state are known, the rest is merely a matter of mathematics. Newton's equations of motion plus the initial data suffice to uniquely solve the mechanical problem. In principle, the future motion of the particle as well as the past motion can be established for any desired period of time. This concept amazed scientists at one time. Laplace, the great French scientist and thinker, once said: If we knew the initial coordinates and velocities of all the particles comprising the world, we would be able to predict the fate of the world. This somewhat naive viewpoint, reducing all reality to purely mechanical phenomena, is not valid in principle, and not merely because it is practically impossible to obtain the required data. Mechanics, based on Newton's laws, has limited application and its conclusions cannot be applied that broadly.

Let us return, however, to the six initial conditions. The need for giving precisely six quantities for a particle is evident from Newton's equations themselves.

The vector equation can be resolved into its components and written in the form of three equations:  $ma_x = F_x$ ,  $ma_y = F_y$  and  $ma_z = F_z$ . To determine the motion, it is necessary to establish how the particle's three coordinates  $x$ ,  $y$ ,  $z$  vary with time. In order to establish the dependence of the coordinate  $x$  on time, we must integrate the equation

$$m \frac{dv_x}{dt} = F_x.$$

The first integration enables us to find the  $x$ -component of the velocity. Upon integrating, we obtain the first constant of integration. The second integration enables us to find the coordinate  $x$  as a function of time, and the second arbitrary constant is obtained. The above also holds true for the equations of change with respect to time for the other two coordinates. In all, six arbitrary constants are obtained. These may be determined only if six independent facts about the coordinates and velocities of the particle are known.

As we have indicated, the initial conditions consist of the three initial coordinates and the three projections of the initial velocity. However, the problem could also be solved if six other quantities are known. For example, we may be given the three coordinates of the initial point, the numerical value of the initial velocity, and two coordinates of the final point. The path of the particle is also uniquely determined by these six conditions.

The parameters of the particle may be given in a variety of ways. The location of the particle in space may be given by three Cartesian coordinates or by the distance from the origin of the coordinate system and two angles formed by the radius vector with the axes. Similarly for the velocity.

A typical example of the dependence of a body's motion on the initial conditions is the behaviour of a rocket fired from the surface of the Earth. The trajectory of the rocket and its destiny is determined by the firing direction, the geographical location of the launching site and the magnitude of the initial velocity. As is well known, for small firing velocities from the Earth, a body has a parabolic trajectory. For a velocity of about 8 km/sec, equilibrium is achieved between the centrifugal force and the gravitational force, and the launched body may be placed in a circular orbit. For velocities between 8 and 11.2 km/sec, the launched body moves in an elliptical orbit about the Earth. At an initial velocity of about

11.2 km/sec, the kinetic energy of the body becomes sufficient to completely overcome the Earth's gravitational attraction. A rocket launched with such a velocity will have a hyperbolic trajectory.

If the mechanical system consists of  $n$  independent points, the number of parameters for the system will be equal to  $6n$ .

In some cases, however, constraints which serve to decrease this number may be placed on the mechanical system. A simple example is a centrifugal regulator, which may be considered as a system consisting of two joined spheres that can slide apart and turn about a common axis. It is clear that, given the distance of a point from the axis of rotation and the azimuthal angle with respect to an arbitrary line, we can uniquely determine the mechanical state of the system. Two "coordinates" and two velocities of change of these coordinates constitute the parameters of this state.

Let us now consider an arbitrarily rotating solid body and determine the data required to fix its position with respect to a stationary system of coordinates. It is clear that the centre of mass of the body is determined by three quantities. To describe the body's rotation, three angles suffice. We need not elaborate on this point, for it is evident that by means of three rotations about mutually perpendicular axes any desired orientation of a body can be achieved.

Thus, the solid body requires twelve parameters—six coordinates and six velocities of change of these coordinates.

As another example, let us consider two rigidly joined points. If they were free, six coordinates would be required to describe them. Since they are rigidly joined, an additional condition relating the coordinates of these points exists, namely:

$$(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2 = \text{const.}$$

Thus, five independent quantities are required to describe this system. In all, there are ten parameters—five coordinates and five velocities of change of these coordinates.

Since the parameters of state are always equally divided between "coordinates" and velocities of change of the "coordinates", it is customary to speak of the *degrees of freedom* of a system, whereby we mean the number of independent coordinates required to describe the system. Thus, one point has three degrees of freedom, two rigidly joined points have five degrees of freedom, a solid body—six degrees of freedom, a system consisting of  $n$  independent points— $3n$  degrees of freedom, etc. The meaning of the following proposition should now be clear: The mechanical state of a system is described by giving its parameters in terms of the number of degrees of freedom.

## Sec. 8. CONSTANTS OF PROPORTIONALITY AND DIMENSIONS OF PHYSICAL QUANTITIES

The coefficient  $\gamma$  in the expression for the law of universal gravitation is a universal constant depending on the choice of units for force, mass and distance. It is possible to choose the units in such a manner that  $\gamma = 1$ . This would require that the unit of mass be equal to the mass of a particle attracting a similar mass at unit distance with unit force. In the CGS system of units, such a mass would be equal to  $1.5 \times 10^7$  g, i.e., 15 tons.

Thus, universal constants in formulas of physics depend on the specific choice of units. If we desired, we could eliminate all such constants from formulas of physics by appropriately choosing the units.

It is important to grasp the concept that the employed system of units and the constants of proportionality in formulas are interconnected. We can demonstrate this interconnection by dimensional formulas. First, the number of units that we wish to consider fundamental must be established. This number depends entirely on us and is determined exclusively by considerations of convenience.

A widely used system of units in physics is based on the units of length ( $L$ ), mass ( $M$ ) and time ( $T$ ) as the independent quantities. The values of all universal constants and the units of measurement of all other quantities are then uniquely determined by the choice of units for  $L$ ,  $M$  and  $T$ . The nature of this relationship is given by so-called dimensional formulas. Several examples will make their meaning clear. The dimensions of velocity are  $LT^{-1}$ , acceleration— $LT^{-2}$ , force— $MLT^{-2}$ , the gravitational constant— $M^{-1}L^3T^{-2}$ , electric charge in the formula for Coulomb's law— $M^{1/2}L^{3/2}T^{-1}$ , etc. Knowing these formulas, we can immediately say how the numerical values of the universal constants and the units of derived physical quantities vary when the magnitude of some fundamental quantity is changed.

As we shall see by examples in Sec. 81, dimensional analysis of physical quantities can be used to predict the nature of some dependence or other between physical quantities.

In addition to the system based on distance, time and mass, a system in which the fundamental quantities are distance ( $L$ ), time ( $T$ ) and force ( $F$ ) is widely used. This is known as the FLT system. Naturally, the dimensional formulas in this system will not always be the same as above. Thus, moment of force in the FLT system has the dimensions  $FL$ , and in the MLT system the dimensions  $ML^2T^{-2}$ . Mass, being a derived quantity in the FLT system, has the dimensions  $FL^{-1}T^2$ .

The fundamental law of mechanics relates the quantities of force, mass, distance and time. Therefore, the value of the constant of proportionality in this formula depends in both systems on the choice of units. In both systems, a constant of proportionality equal to unity is assumed. This means that in the MLT system, using the formula  $F = ma$ , the unit of force is chosen so that  $F = 1$  when the mass and the acceleration are equal to unity. In the FLT system, using the formula  $m = \frac{F}{a}$ , the unit of mass is chosen so that  $m = 1$  when the force and the acceleration are equal to unity.

In this book, we shall for the most part use two variants of the MLT system:

CGS system: L—centimetre, M—gram, T—second;

SI system: L—metre, M—kilogram, T—second.

In the CGS system, the unit of force is the dyne  $= 1 \text{ g} \times \text{cm/sec}^2$  and the unit of work is the erg  $= \text{dyne} \times \text{cm}$ . In the SI system, the unit of force is the newton  $= 1 \text{ kg} \times \text{m/sec}^2$  and the unit of work is the joule  $= \text{newton} \times \text{metre}$ .

If the reader is confronted with data expressed in the FLT system, these data should be converted into one of the indicated systems. To do this, he only need recall that a unit of force in the FLT system is a kilogram-force (the weight of a kilogram mass at sea level at  $45^\circ$  latitude), which is related to the two units of force adopted by us as follows:

$$1 \text{ kgf} = 9.81 \text{ newtons} = 9.81 \times 10^5 \text{ dynes.}$$

We shall return to the subject of systems of units when we consider electrical quantities.

# Mechanical Energy

## Sec. 9. WORK

Motion without acceleration (i.e., rectilinear and uniform) may take place either without the action or with the action of forces on the body. In the latter case, the sum of the forces acting on the body is equal to zero. There is an essential difference between these two kinds of motion. In the first case, motion is not accompanied by work, while to achieve the second type of motion work must be expended. A motor works, moving an automobile uniformly and rectilinearly. A man works, moving a sleigh with its load uniformly and rectilinearly. We say in these cases that work is expended in overcoming resistance—friction, air resistance, etc.

Of the two balanced forces acting on a body moving without acceleration, one is directed along and the other opposite to the direction of motion.

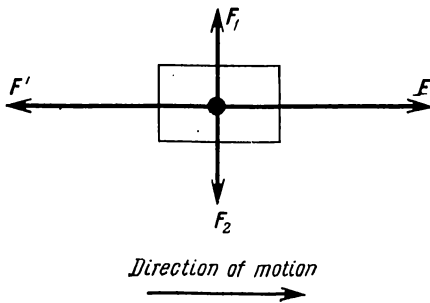


Fig. 15

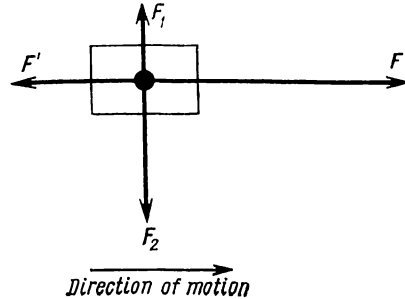


Fig. 16

We say that a force acting along the direction of motion performs work. On the other hand, as regards a force directed opposite to the motion, we say that work is performed against this force.

For quantitative evaluation, work is expressed as the product of the force acting on a body and the distance traversed by the body. The term *work* is the designation for this physical quantity.

Let a body be acted upon by a number of forces whose vector sum is equal to zero. The body moves uniformly and rectilinearly. All the forces may then be resolved into four components (Fig. 15). Forces  $F_1$  and  $F_2$  in accordance with the adopted definition, perform no work. Force  $F$  performs work equal to  $F\Delta S$ , where  $\Delta S$  is the traversed distance. The work of force  $F'$  is equal to  $-F'\Delta S$ , where the minus sign indicates that work is performed against the force  $F'$ .

Let us now consider the motion of a body with acceleration, i.e., curvilinear and nonuniform motion. As we already know, a resultant force acts on the body, in this case, that is directed along the acceleration (but not along the path in the general case!). Let us again resolve all the effective forces into forces directed along the motion and perpendicular to the motion (Fig. 16). Now,  $F$  is not equal to  $F'$ , and  $F_1$  is not equal to  $F_2$ . Using the definition for work given above, we can assert, as before, that  $F_1$  and  $F_2$  perform no work. The work of force  $F'$  is again negative, i.e., the work is performed against the force  $F'$  and is equal to  $-F'\Delta S$ .

Force  $F$  performs the work  $F\Delta S$ , which is more than the work against the force of resistance. The surplus work serves to accelerate the body.

The inequality between the forces  $F_1$  and  $F_2$  shows that the motion is curvilinear. Their difference  $F_2 - F_1$  corresponds to the normal component of the acceleration.

Let us consider an extreme case—uniform motion in a circle. The resultant force for such motion is directed, as we know, along the radius of the circle, i.e., perpendicular to the direction of motion. Therefore, the centripetal force performs no work.

Thus, the surplus work in the general case of curvilinear accelerated motion is not used to produce acceleration in general, but only the tangential component of the acceleration. For a particle, this can be expressed as follows:

$$F - F' = ma_t \quad \text{and} \quad F\Delta S - F'\Delta S = ma_t \Delta S.$$

We repeat:  $(F - F')$  is the tangential component of the resultant force  $F_t^{res}$ .

The work expended in accelerating a body (being equal, by definition, to the projection of the resultant force on the direction of motion multiplied by the traversed distance) is equal to the product of the mass of the body, the traversed distance, and the tangential acceleration. The last equation above can be written in the form  $F\Delta S = F'\Delta S + ma_t \Delta S$  and can be read as follows: The work performed by the effective force consists of the work against the force of resistance and the work expended in accelerating the body.

*Examples. 1.* A jet passenger plane, having a weight  $P = 70$  tf, attains a height  $h = 10$  km. If it moves uniformly, the work performed in rising to this height is

$$A_1 = Ph = 7 \times 10^8 \text{ kgf-m} = 68.6 \times 10^8 \text{ joules} = 68.6 \times 10^{15} \text{ ergs}.$$

If this height is attained over a path  $S = 85$  km with a simultaneous increase in velocity (acceleration  $a = 0.3$  m/sec<sup>2</sup>), the additional expenditure of work in producing acceleration will be

$$A_2 = ma \times S = 17.9 \times 10^8 \text{ J} = 17.9 \times 10^{15} \text{ ergs} = 1.82 \times 10^8 \text{ kgf-m}.$$

2. To plane a board 2 metres long and 20 centimetres wide, a joiner expends about 150 kgf-m = 1470 J of work.

## Sec. 10. KINETIC ENERGY

Thus, in accelerating a body, the resultant force  $F^{res}$  performs the work

$$A = F_t^{res} \times \Delta S = ma_t \Delta S,$$

where  $a_t$  is the average tangential acceleration over the portion of path  $\Delta S$  being considered. Substituting for  $a_t$ , we obtain

$$A = m \frac{\Delta v \times \Delta S}{\Delta t} = mv \times \Delta v,$$

where  $v$  is the average velocity which is equal to  $(1/2)(v_2 + v_1)$ . Here  $v_1$  and  $v_2$  are the instantaneous velocities at the beginning and the end of the path respectively. Since  $\Delta v = v_2 - v_1$ , then\*

$$A = \frac{mv_2^2}{2} - \frac{mv_1^2}{2} = \Delta \left( \frac{mv^2}{2} \right),$$

---

\* The same result is obtained if we write the expression for an infinitely small quantity of work in the form  $dA = mv dv$  and integrate it from the moment the velocity was  $v_1$  to the moment it reaches  $v_2$ :

$$A = \int_{v_1}^{v_2} mv dv = \frac{mv_2^2}{2} - \frac{mv_1^2}{2}.$$

i.e., the work is numerically equal to the increment in the value of  $\frac{mv^2}{2}$ . Therefore, the quantity

$$K = \frac{mv^2}{2}$$

is employed as a measure of the energy of motion of a particle. This quantity  $K$  will be called *kinetic energy*. The previous equation may now be read as follows: the work of the resultant force acting on a body (i.e., the product of the tangential component of the resultant force and the traversed distance) is equal to the increment in the kinetic energy of the body. This equation is convenient for the solution of elementary mechanical problems in which the path along which the force acts is given.

We shall repeatedly be dealing with the term "energy". It is one of the most important physical concepts. Energy, i.e., work capacity, is a function of the state of a body. Work is produced at the expense of a decrease in the value of this function. Kinetic energy is a function of the state of motion. If the kinetic energy changes from  $K_1$  to  $K_2$ , then the work performed thereby is equal to  $K_2 - K_1$ , independent of the nature of the motion. It is of no importance whether the velocity changed rapidly or slowly, uniformly or nonuniformly. The decrease in the kinetic energy by a specific amount always yields the same amount of work.

Only in the case when the physical quantity is a function of state can it have the sense of energy, i.e., a store of work.

*Examples.* A unit of energy in atomic physics is the electron-volt (eV). This is the kinetic energy of an electron accelerated through a potential difference of 1 volt:

$$1 \text{ eV} = 1.6 \times 10^{-12} \text{ erg} = 1.6 \times 10^{-19} \text{ J}.$$

The energy of a proton accelerated in a synchrotron is  $10 \text{ GeV} = 10^{10} \text{ eV} = 0.016 \text{ erg} = 1.6 \times 10^{-5} \text{ J}$ .

The kinetic energy of a large jet passenger plane ( $m = 100$  tons and  $v = 800$  km/hr) is

$$2.5 \times 10^{16} \text{ ergs} = 2.6 \times 10^9 \text{ J} = 2.5 \times 10^8 \text{ kgf-m}.$$

## Sec. 11. POTENTIAL ENERGY

Let us consider several phenomena in which the performed work is not accompanied by a change in the velocity of the body. We shall be concerned with two types of problems. The first is related to the elastic deformation of a body, while the second deals with events occurring during the motion of a body in a gravitational or electric field. We shall presently show that in both cases we shall be dealing with the transformation of work into a special variety of energy, so-called potential energy.

Elastic deformation phenomena will be treated first. Experiments show that for any elastic deformation—extension, compression, flexure, etc.—one can always find a function of state that increases precisely by the magnitude of the work performed on the body. This function of state or, in other words, the function of the body's properties and degree of deformation is called *the potential energy of elasticity*.

We shall show this energy to exist for a case of elastic deformation, namely, linear extension or compression. Analogous examples could be given for any other kind of elastic deformation.

Let some force, such as a muscular force, stretch a solid body, e.g., a spring, very slowly. The work expended in stretching the body from length  $l + s_1$  to

length  $l + s_2$ , where  $l$  is the length of the unstretched spring, is

$$A = F(s_2 - s_1).$$

The muscular force is balanced by the elastic force of the spring at every given moment. For a small extension of the spring, the latter is proportional to the deformations  $s^*$ :

$$F_{el} = ks.$$

In the expression for work, we must use the average value of the force  $F$ , i.e.,  $\frac{1}{2}(ks_1 + ks_2)$ . We then obtain\*\*:

$$A = \frac{ks_2^2}{2} - \frac{ks_1^2}{2} = \Delta \left( \frac{ks^2}{2} \right),$$

i.e., the work against the elastic force is expended in increasing the quantity  $\frac{ks^2}{2}$ . This quantity is, therefore, adopted as a measure of the elastic energy. The quantity

$$U_{el} = \frac{ks^2}{2}$$

will be called *elastic potential energy*.

Elastic potential energy formulas for other kinds of deformation have exactly the same form. The body's stiffness with respect to a specific form of deformation is characterised by  $k$ , while  $s$  is a measure of the deformation (for example, twist angle, displacement angle, etc.).

The quantity  $U_{el}$  is energy in precisely the sense referred to at the end of Sec. 10. Irrespective of the manner in which a body is deformed and the rapidity of the process, the same amount of expended work will always correspond to one and the same incremental value of the quantity  $\frac{ks^2}{2}$ . Thus,  $\frac{ks^2}{2}$  is a measure of energy or, to be more precise, of elastic potential energy.

*Examples.* 1. The potential energy of a piece of steel wire (Young's modulus  $E = 20.6 \times 10^{10}$  N/m<sup>2</sup>) having a length of 50 metres and a cross-section of 10 mm<sup>2</sup>, and which is stretched 1 cm, is

$$U_{el} = \frac{ks^2}{2} = 20 \times 10^6 \text{ ergs} \approx 2 \text{ J}.$$

2. For rubber, Young's modulus  $E = 7.85 \times 10^5$  N/m<sup>2</sup>. A stone having a mass of 20 g is shot from a slingshot to a height of 20 metres. This requires that an energy of 3.92 J be imparted to the stone. Assume that the elastic band has an initial length of 40 cm and stretches an additional 40 cm. Let us determine the required cross-section for the elastic band.

$$U_{el} = \frac{ES}{l} \frac{s^2}{2}; \quad S = \frac{2LU}{Es^2} = \frac{2 \times 40 \text{ (cm)} \times 3.92 \text{ J}}{7.85 \times 10^5 \text{ N/m}^2 \times 1,600 \text{ (cm)}^2} = 0.25 \text{ cm}^2.$$

Gravitational force possesses the same feature as elastic force. Thus, work expended in lifting a body in a gravitational field serves to change the body's function of state. In this case, the function interesting us depends on the position of the

---

\* It should be recalled that the law of elastic deformation (Hooke's law) is written in the form  $\frac{F}{S} = E \frac{s}{l}$ , where  $E$  is the modulus of elasticity and  $S$  is the cross-section of the stretched body. Thus, the stiffness (the constant of proportionality in the expression for the elastic force) has the value  $k = \frac{ES}{l}$ .

\*\* The same result is obtained when we integrate the infinitely small amount of work  $dA = -ks ds$  between the limits  $s_1$  and  $s_2$ .

given body with respect to the bodies attracting it. This function is called *gravitational potential energy*.

We shall show this energy to exist, first, for a body located close to the Earth's surface. From point 1, the body is moved to the higher point 2 along some curvilinear path. Let us divide this path into small segments, replacing the curved line by a broken line. The latter can be made to approximate the former to any desired accuracy. The work expended in moving a body along one of these linear segments of length  $dl$  is then

$$dA = mg \, dl \sin \alpha \quad \text{or} \quad dA = mg \, dh,$$

where  $dh$  is the increase in height. Since  $mg$  does not change along the entire path of motion, we can place it before the brackets (before the integral sign when integrating) in writing the expression for the work expended along the entire path:

$$A = mg(h_2 - h_1),$$

where  $h_1$  and  $h_2$  are the heights of points 1 and 2 respectively. Furthermore,

$$A = (mgh)_2 - (mgh)_1 = \Delta(mgh),$$

i.e., the work of displacement is equal to the increase in the product  $mgh$ , which is a measure of the gravitational potential energy for this simple case.

It is quite evident that

$$U = mgh$$

is energy and is in complete accord with the meaning we have assigned to this term. Irrespective of the manner in which the work is performed, i.e., the path taken by a body and the speed of the motion, the work of displacing the body from point 1 to point 2 will always be the same, since the increase of energy depends only on the location of these points—in our simple case, on their heights.

Since the work of displacing a body in a gravitational field does not depend on the path, the work of displacement along a closed curve will be equal to zero.

It should be noted that it is immaterial what level we choose as our base for  $h$ . If it is agreed to calculate  $h$  from the Earth's surface, the potential energy of a body at the bottom of a well will be negative.

The above formula is not valid for bodies that are far from the Earth, e.g., the Moon. Thus, as was explained in Sec. 2,  $mg$ , the approximate expression for the gravitational force, should be replaced for large distances by the exact expression  $\gamma \frac{m_1 m_2}{r^2}$ .

Let us calculate the work done by the gravitational force. Work performed by the forces of a system will be considered positive, while work against the forces of the system will be considered negative. Let us assume that two attracting bodies draw together along the line of action of the forces over an infinitely small segment  $-dr$  of the path (minus, since  $r$  decreases). Thus,

$$dA = -\gamma \frac{m_1 m_2}{r^2} dr.$$

But  $\frac{dr}{r^2} = d\left(-\frac{1}{r}\right)$ . Therefore,

$$dA = -d\left(-\gamma \frac{m_1 m_2}{r}\right).$$

Work takes place at the expense of a decrease in the value of  $U = -\gamma \frac{m_1 m_2}{r}$ , which is a measure of the gravitational energy in the general case:

$$dA = -dU.$$



The quantity

$$U = -\gamma \frac{m_1 m_2}{r}$$

represents gravitational potential energy in the general case.

$U$  is equal to zero if the bodies are infinitely far apart. When the bodies draw together,  $U$  increases in absolute value. But since  $U$  is negative, we see that, just as with the approximate formula for bodies close to the Earth, the potential energy is less the closer the attracting bodies are to each other. Naturally, if we desired, we could change the base line for  $U$  and make this quantity positive in the interval of values concerning us.

It is not difficult to show the relationship between the general formula for  $U$  and its particular case when  $U = mgh$ . Thus, replacing  $r$  by  $R + h$ , where  $R$  is the radius of the Earth, we obtain

$$U = -\gamma \frac{Mm}{R+h} = -\frac{\frac{1}{R} \gamma Mm}{1 + \frac{h}{R}}$$

( $M$  is the mass of the Earth). But since  $\frac{h}{R}$  is a small quantity, we can write with sufficient accuracy  $\frac{1}{1 + \frac{h}{R}} = 1 - \frac{h}{R}$ , whence

$$U = -\gamma \frac{Mm}{R} + mgh.$$

Changing the base line for  $U$  so that zero potential energy for a body is at the Earth's surface, the formula reduces to  $U = mgh$ .

*Example.* To obtain a clearer picture of the meaning of the above results, let us calculate the potential energy of a body of mass  $m = 1$  kg at the Earth's surface and at a distance of 1,000 km above its surface.

The potential energy at the surface of the Earth is

$$U_0 = -\gamma \frac{Mm}{R} \approx 6.67 \times 10^{-11} \times \frac{5.8 \times 10^{24} \times 1}{6.3 \times 10^6} = -6.1 \times 10^7 \text{ J} = -6.1 \times 10^{14} \text{ ergs.}$$

The potential energy at a height of 1,000 km is

$$U_{1,000} = -6.67 \times 10^{-11} \times \frac{5.8 \times 10^{24} \times 1}{7.3 \times 10^6} = -5.3 \times 10^7 \text{ J} = -5.3 \times 10^{14} \text{ ergs.}$$

From the calculations, it is evident that 1) the potential energy of a body in the Earth's gravitational field is always negative and increases with its distance from the Earth (since we have agreed that it tends to zero when  $h \rightarrow \infty$ ); 2) the change in the potential energy of a body rising above the Earth's surface is, generally speaking, not described by the expression  $mg(h_2 - h_1)$ . Thus,

$$U_{1,000} - U_0 = -5.3 \times 10^7 - (-6.1 \times 10^7) = 0.8 \times 10^7 \text{ J}$$

while the calculation with the expression  $mg(h_2 - h_1)$  yields  $0.98 \times 10^7 \text{ J}$ . However, when we are concerned with ascensions to a height  $h \ll R$  ( $R$  is the radius of the Earth), it is permissible to use the simplified expression  $mg(h_2 - h_1)$ .

The formula for the *potential energy of the electrical interaction of charges* is very similar to that for gravitational potential energy.

Let us consider two electrical charges  $q_1$  and  $q_2$  having the same sign and separated by a distance  $r$ . According to Coulomb's law, the particles will repel each other. Therefore, in reducing their separation to the small distance  $dr$ , we perform

work equal to  $-dA = -\frac{q_1q_2}{r^2} dr$  (the minus sign is used in the left-hand member because the work is performed against the forces of the system; the right-hand member also has a minus sign because the distance is decreasing and  $dr$  is negative). The calculation, which in no way differs from that for gravitational force, yields for the energy of electrical interaction of charges (for brevity called Coulomb energy) the expression  $U = \frac{q_1q_2}{r}$ , i.e., here too  $dA = -dU$ .

The interaction energy of charges having opposite signs is negative and behaves like gravitational energy. The interaction energy of charges having the same sign is equal to zero when the charges are separated by an infinite distance; it increases as the charges are brought together.

We shall restrict ourselves to these examples of potential energy, although in various cases other functions of state of a body may be introduced.

Potential energy always exists when forces act between bodies or particles of the system under consideration that depend on the distance between the bodies. Potential energy is the interaction energy of the bodies. If a system consists of a number of bodies or particles, we can then speak of its total potential energy, i.e., the interaction energy between all the particles (each with all the rest). Thus, in the case of four particles, the potential energy is composed of six terms, for we must consider the interaction between the first body and the second, third and fourth; the second body and the third and fourth; and finally the third body and the fourth.

In mechanics, only the potential energy of forces acting between different bodies is considered. If a body is complex and consists of many particles, the interaction potential energy of these particles is considered to remain unchanged during the mechanical processes. The interaction potential energy of the particles comprising the body is a component part of the body's internal energy (Chapter IX). If changes in a body's internal energy take place, the phenomena must be considered in the light of the laws of thermodynamics (Chapter IX).

## Sec. 12. LAW OF CONSERVATION OF MECHANICAL ENERGY

Irrespective of the type of forces involved in the motion, the work of the resultant force is always equal to the increment of the body's kinetic energy, i.e.,

$$\sum \vec{F} \Delta s = \Delta \left( \frac{mv^2}{2} \right).$$

The forces acting on the body could be elastic forces, gravitational, electrical and frictional forces, etc.

It is always possible to separate from the effective forces those whose work serves to change the potential energy. For brevity, such forces are sometimes called *potential* forces or forces possessing potential. The work equation may be written in the form

$$F_{pot} \Delta s + f \Delta s = \Delta \left( \frac{mv^2}{2} \right).$$

Here,  $f$  represents the nonpotential forces. The work of these forces is equal to the change in the internal energy of a body or the medium in which the body moves.

Substituting in place of the work of the potential forces the increment of potential energy with reversed sign, we can write the equation in the form

$$f\Delta s = \Delta \left( \frac{mv^2}{2} + U \right).$$

The sum of a body's potential and kinetic energy is called *total mechanical energy*. Designating this quantity by  $\mathcal{E}$ , we obtain:  $f\Delta s = \Delta\mathcal{E}$ , i.e., the change in a body's total energy is equal to the work of the nonpotential forces, e.g., the frictional forces.

If the work performed in changing the body's internal energy is small with respect to  $\mathcal{E}$ , the equation simply reduces to the following:  $\Delta\mathcal{E} = 0$  and  $\mathcal{E} = \text{const}$ . This is the law of conservation of mechanical energy, which states that the total mechanical energy of a body is conserved.

This law may be easily generalised for a system consisting of many bodies or particles. For each body we may write a work equation and then combine these equations into one. The total energy will then be equal to the sum of the kinetic energies of the bodies and the potential energy of interaction:

$$\mathcal{E} = \frac{m_1 v_1^2}{2} + \frac{m_2 v_2^2}{2} + \dots + U.$$

If all the interacting bodies are taken into account (such a system of bodies is called a *closed system*), the form of the law remains the same as for a single body. The change in mechanical energy is equal to the work of the nonpotential forces and, if this work is negligible, the total mechanical energy of a closed system of bodies remains unchanged, i.e., is conserved.

The law of conservation of mechanical energy is, on the one hand, a consequence of the equations of mechanics (Newton's law); on the other hand, it may be considered as a special case of a more general law of nature—the law of conservation of energy (Chapter IX).

Even in mechanics alone, many forms of interconvertible energy are met. In considering the motion of a body under the action of elastic forces or gravitational force, it is easily seen that an increase in the energy of one of the mechanical forms is accompanied by a decrease in the energy of the other form.

Thus, the gravitational force acting on a falling body decreases the potential energy of the body and increases its kinetic energy. The reverse is true when a body is lifted to a certain height. The elastic force making a ball thrown against a wall rebound decreases the potential energy of the compressed ball and transforms it into kinetic energy. The reverse takes place when the wall stops the thrown ball (the interval from no deformation to maximum compression).

A stretched spring can raise a load to a certain height. On the other hand, a falling load can stretch a spring. Thus, elastic energy can be transformed into gravitational energy and vice versa.

The above examples apply to the transformation of one form of energy into another in one and the same body as well as to the transfer of energy from one body to another.

It is possible, of course, to transfer energy in the same form from one body to another: one load pulls another by means of a pulley, a sphere colliding with another transfers part of its kinetic energy to it, etc.

## Sec. 13. POTENTIAL CURVES. EQUILIBRIUM

The potential energy of interaction of bodies or particles depends on their relative distribution, i.e., it is always a function of the coordinates or other parameters describing the location of these bodies in space. In the simplest cases, the potential energy may depend on a single coordinate.

Let us consider the interaction of two particles whose potential energy of interaction is described by the function  $U(x)$ , where  $x$  is the distance between the particles. For the sake of definiteness, let us assume the particles repel each other with a force  $F$ . Under the action of the interaction forces, the distance between them increases by  $dx$ , i.e., an amount of work equal to  $Fdx$  is performed. This is possible at the expense of the potential energy of interaction  $U$ , which changes by  $-dU$  (decrease of energy).

Thus,  $-dU = F dx$   
or

$$F = -\frac{dU}{dx},$$

i.e., in the case of potential forces the force is equal to minus the derivative of the potential energy with respect to  $x$ . The nature of the mechanical problem is then very simple and is clearly described by so-called potential curves, i.e., graphs on which the values of the potential energy are plotted as a function of the parameter (Fig. 17).

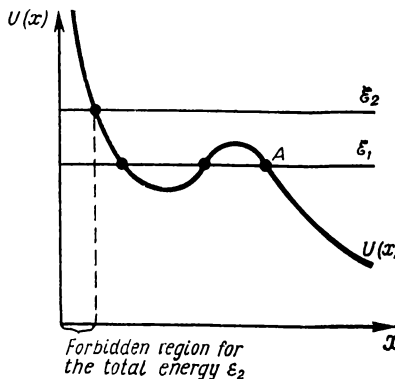


Fig. 17

In explaining the essence of this graphical method, the analogy is usually drawn with the motion of a body on a hill. The meaning of the potential curve then becomes particularly clear, for the profile of a hill and a potential energy distribution curve that is proportional to the height  $h$  of the hill coincide if drawn to proper scale. ■

Potential curves consist of crests and troughs, steep and gradual rising slopes as well as steep and gradual declining slopes. The form of the curve permits us to immediately indicate on

which portions of the path a large amount of work is performed, on which a small amount, and whether the work is positive or negative in each case. The steeper the potential curve, the larger the force acting on the body. In accordance with the familiar geometric sense of the derivative, force is described by the tangent of the angle of inclination of the tangential line to the potential curve.

The validity of the formula relating potential energy and force is completely evident for those particular cases of potential energy that we have considered. For the potential energy of a body on the Earth's surface:

$$U = mgh \quad \text{and} \quad F = -\frac{dU}{dh} = -mg.$$

For a body in a gravitational field, in the general case:

$$U = -\gamma \frac{m_1 m_2}{r} \quad \text{and} \quad F = -\frac{dU}{dr} = -\gamma \frac{m_1 m_2}{r^2}.$$

For a body subjected to elastic action:

$$U = \frac{kx^2}{2} \quad \text{and} \quad F = -\frac{dU}{dx} = -kx$$

For electrical interaction:

$$U = -\frac{q_1 q_2}{r} \quad \text{and} \quad F = -\frac{dU}{dr} = -\frac{q_1 q_2}{r^2}.$$

Returning to the potential curve plotted in the diagram and keeping the above explanation in mind, we can immediately indicate where the force is greatest and the points where the force acting on the body is equal to zero. The latter points, i.e., the positions of equilibrium, are at the bottom of the potential well and at the potential peak. The positions where the potential energy is a maximum correspond to unstable equilibrium, while the bottom of the potential well is a position of stable equilibrium.

We stated above that the form of the potential curve permits us to describe the possible motion of the body. This is not completely accurate. In addition

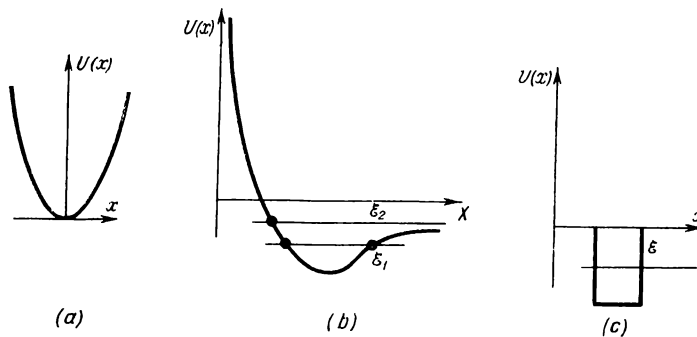


Fig. 18

to the potential curve, we must also know the value of the total mechanical energy of the body. If this value is known, we can then indeed deduce from the form of the potential curve the possible motion of the body or particle.

Horizontal lines are drawn in Fig. 17 at the ordinates corresponding to  $\mathcal{E}_1$  and  $\mathcal{E}_2$ . If  $\mathcal{E}$  is the total energy of the particle, then we can determine the kinetic energy as well as the potential energy from the curve. The former is the difference between  $\mathcal{E}$  and  $U$ .

The moving particle cannot occupy positions in which the potential energy is greater than the total energy. Thus, the horizontal line  $\mathcal{E}$  restricts the possible motion of the body to certain portions of the curve. In the case when the energy is represented by the lower line  $\mathcal{E}_1$ , the moving point has two possible intervals in which it may be located. It may be either in the potential well (and have an oscillatory motion there) or on the slope to the right of point A, where it will move downwards or upwards depending on whether it acquires or loses kinetic energy.

The above analysis is completely valid for any kind of potential curve. Fig. 18 shows several types of such curves. Thus, in Fig. 18a, we see the potential curve for a body oscillating on a spring. The oscillating body is in a potential well having symmetrical edges. In Fig. 18b, a potential curve is shown that is typical for many interacting particles, e.g., atoms and molecules. The curve constitutes a potential well, one of whose edges has a very steep slope while the other has a gradual one. Plotted along the abscissa is the distance between the particles. As can be seen from the curve, the potential energy is very large at small distances, falls with increasing distance, reaches a minimum, then gradually rises tending

toward some finite limit. The nature of the motion and the bond between two interacting particles are completely described by this curve. Two cases should be distinguished. The first is when the total mechanical energy of this pair of particles is represented by the lower horizontal line  $\mathcal{E}_1$ . The second is when the total energy is equal to  $\mathcal{E}_2$ . In the first case, the system cannot get out of the potential well. This means that the distance between the particles lies between the limits indicated in the figure. The mutual motion of the particles can only be of an oscillatory character. Such is the situation in a stable diatomic molecule. The second case is the reverse of the first. The total energy of interaction of the particles is too large for them to be constantly linked. The system may get out of the potential well, i.e., the bond between the particles cannot exist and the particles may fly apart to any distance whatsoever.

The third potential curve in the figure is a so-called square well. Recalling that force is described by the tangent of the angle of inclination of the tangential line to the potential curve, we see that the potential energy may be represented in the form of a square well if the body or particle moves freely without the action of forces, yet cannot leave the bounds of the given portion of the curve as long as the total energy is less than the height of the sides of the well.

# Momentum

## Sec. 14. CONSERVATION OF MOMENTUM

The product of the mass of a body, or particle, and its velocity is known as the body's momentum (quantity of motion):  $\mathbf{p} = m\mathbf{v}$ . The momentum  $\mathbf{p}$  is thus a vector quantity. In a system of bodies or particles, the momentum is equal to the vector sum of the particles constituting the system:

$$\mathbf{P} = \mathbf{p}_1 + \mathbf{p}_2 + \dots$$

What makes this vector quantity of particular interest to the physicist is the fact that in a closed system the vector  $\mathbf{P}$  does not change, irrespective of the motion within the system itself. This proposition is known as the *law of conservation of momentum*.

The law of conservation of momentum follows directly from Newton's laws. For each of the bodies in the closed system, the following equation is valid:

$$\frac{d}{dt}(m\mathbf{v}) = \mathbf{F},$$

i.e.,

$$\frac{d\mathbf{p}}{dt} = \mathbf{F}.$$

Let us consider what happens when we write such an equation for each of the bodies and then add the equations. The right-hand member of each equation represents the forces exerted on the given body by all the other bodies. Thus, the force exerted on the first body is equal to the sum of the forces exerted on it by the second, third, etc. Using double indexes, we may write:  $\mathbf{F}_{12} + \mathbf{F}_{13} + \mathbf{F}_{14} + \dots$ . Similarly, for the forces exerted on the second body, we may write:  $\mathbf{F}_{21} + \mathbf{F}_{22} + \mathbf{F}_{23} + \dots$ ; for the third:  $\mathbf{F}_{31} + \mathbf{F}_{32} + \mathbf{F}_{33} + \dots$ ; etc. It is not difficult to see that when the right-hand members of the equations are added the result is zero. For each term in the first line, there is always a term in another line that is equal and opposite to it (in accordance with the law of action and reaction). Thus, when  $\mathbf{F}_{12}$  and  $\mathbf{F}_{21}$  are added the result is zero; also  $\mathbf{F}_{13}$  and  $\mathbf{F}_{31}$ ; etc. Therefore, in a closed system, the following equation holds:

$$\frac{d\mathbf{p}_1}{dt} + \frac{d\mathbf{p}_2}{dt} + \frac{d\mathbf{p}_3}{dt} + \dots = 0; \quad \frac{d}{dt}(\mathbf{p}_1 + \mathbf{p}_2 + \mathbf{p}_3 + \dots) = 0$$

or

$$\mathbf{p}_1 + \mathbf{p}_2 + \mathbf{p}_3 + \dots = \text{const.}$$

This is the law of conservation of momentum. The magnitude and direction of individual moments may change, but their vector sum in a closed system does not.

*Magnitudes of some momenta:* The momentum of an electron with an energy of 5 eV is  $\sim 12 \times 10^{-20} \frac{\text{g} \times \text{cm}}{\text{sec}}$ ; that of a rifle bullet is  $\sim 8 \times 10^5 \frac{\text{g} \times \text{cm}}{\text{sec}} = 8 \frac{\text{kg} \times \text{m}}{\text{sec}}$ ; and that of a freight train is  $\sim 10^7 \frac{\text{kg} \times \text{m}}{\text{sec}}$ .

## Sec. 15. CENTRE OF MASS

The methods of finding the centre of gravity of a body are well known. If a body is fixed at its centre of gravity, it is in a state of neutral equilibrium. For a system of particles, or a solid body considered to be broken up into elementary elements having the size of particles, we can write an analytical expression for the position of the centre of gravity.

Using the rule for adding parallel forces (Fig. 19), we obtain the following expression for the position of the centre of gravity when the particles are considered to be distributed along a straight line, say the  $x$ -axis:

$$X = \frac{m_1 x_1 + m_2 x_2 + m_3 x_3 + \dots}{m_1 + m_2 + m_3 + \dots}$$

Here,  $x_1, x_2, x_3 \dots$  are the coordinates of the particles, and  $m_1, m_2, m_3 \dots$  are the masses. Masses are used instead of weights since the acceleration of the gravitational force cancels out.

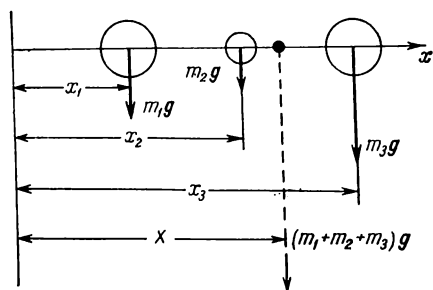


Fig. 19

It is shown in theoretical mechanics that for any distribution of particles the expression for the position of the centre of gravity has the form:

$$\mathbf{R} = \frac{m_1 \mathbf{r}_1 + m_2 \mathbf{r}_2 + m_3 \mathbf{r}_3 + \dots}{m_1 + m_2 + m_3 + \dots},$$

where  $\mathbf{R}$  is the radius vector of the centre and  $\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3$  are the radius vectors of the particles.

Since the acceleration of the gravitational force cancelled out in these formulas, we can conclude that the point found has an objective significance that does not depend on the gravitational conditions. It is valid, in fact, even if the body is located in interplanetary space under conditions of weightlessness. It makes sense, therefore, to replace the prevalent designation "centre of gravity" by a designation that more accurately expresses the essence of the matter. Thus, we speak of the centre of inertia or *centre of mass* of a body instead of its centre of gravity.

We shall directly see the full significance of this designation. Let us consider the velocity of motion of the centre of mass:

$$\mathbf{V} = \frac{d\mathbf{R}}{dt}.$$

Using the formula for the determination of the centre of mass, we obtain:

$$\mathbf{V} = \frac{m_1 \mathbf{v}_1 + m_2 \mathbf{v}_2 + m_3 \mathbf{v}_3 + \dots}{m_1 + m_2 + m_3 + \dots}.$$

In the numerator we have the total momentum, which is conserved in a closed system. Thus, the right-hand member of the equation is equal to a constant quantity. We can conclude, therefore, that the velocity of the centre of mass does not change in magnitude or direction. Or, in other words, the centre of mass of a closed system of particles executes inertial motion.

As we already know, all inertial systems of coordinates have equal validity. Hence, we can always go over to a coordinate system bound to the centre of mass of the system under investigation and consider this interesting point as fixed. In atomic physics, we often consider collisions between particles. To study this



phenomenon, two systems of coordinates are used—the laboratory system (the natural coordinate system of an observer) and the system bound to the centre of mass of the colliding particles. The advantage of the latter frame of reference is evident: the total momentum of the particles is equal to zero.

### Sec. 16. COLLISIONS

The word “collision” should be understood in a somewhat broader sense than that used in everyday practice. For the mechanical problems that now concern us, any encounter between two or more bodies in which the interaction is of short duration will be considered to be a collision. Thus, in addition to the phenomena that can be classified as collisions in the usual sense of the word—e.g., impact of billiard balls and collisions between atoms and atomic nuclei—we have such events as a man jumping on or off a street car and a bullet hitting a wall. The forces arising as the result of such short interactions are so great that the role of all constant forces being exerted is negligible. As a result we are justified in considering the colliding bodies as a closed system and we can apply the law of conservation of momentum to them.

In many collisions, the duration of the interaction is measured in thousandths of a second. During this interval of time, the force rises to its maximum value and then drops to zero. A typical curve for the force during such an impact is shown in Fig. 20. For each instant of time during the impact, the relationship between the force exerted on either of the bodies and the momentum of this body is given by Newton's second law:

$$\frac{d}{dt}(mv) = F.$$

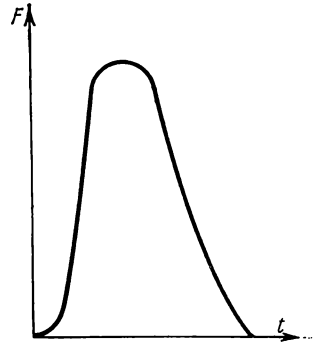


Fig. 20

Rewriting this equation in the form  $F \Delta t = \Delta(mv)$ , we can say that the product of the average value of the force and the duration of its action is equal to the change in momentum. A more accurate description of the phenomenon is obtained if we integrate the above equation from the initial instant of impact to the termination of the interaction. It is evident that

$$\int_0^{\tau} F dt = (mv)_2 - (mv)_1.$$

The integral on the left is sometimes called the impulse of the force. In the diagram, this quantity is represented geometrically by the area under the impact curve (see Fig. 20).

There is considerable variation in the nature of collisions, depending on the elastic properties of the bodies. It is customary to consider two extreme cases—ideally elastic and absolutely nonelastic impacts.

First, let us consider the latter type. A *nonelastic impact* means an encounter between two bodies whereby these two bodies become joined. Examples of nonelastic impacts are collisions between clay spheres, a man jumping onto a moving trolley, the collision between oppositely charged ions resulting in the formation of a molecule, and the capture of an electron by a positive ion.

Assume that the bodies moved with velocities  $v_1$  and  $v_2$  before the encounter. Thus, the total momentum was  $m_1v_1 + m_2v_2$ . After the encounter the bodies have

acquired mass equal to  $m_1 + m_2$  and move with some velocity  $V$ . The momentum of the system after the encounter is  $(m_1 + m_2) V$ . Since the law of conservation of momentum requires that

$$(m_1 + m_2) V = m_1 v_1 + m_2 v_2,$$

the velocity of the bodies after the nonelastic impact is given by the formula:

$$V = \frac{m_1 v_1 + m_2 v_2}{m_1 + m_2}.$$

The momentum after the encounter should equal the sum of the momenta before impact.

If the motion of the bodies colliding head-on is along a straight line, then after impact the bodies will follow the direction of the body having the originally larger momentum. If the momenta of the bodies are equal in magnitude  $m_1 v_1 = -m_2 v_2$ ;  $V$  is thus equal to zero, i.e., the colliding bodies come to a standstill.

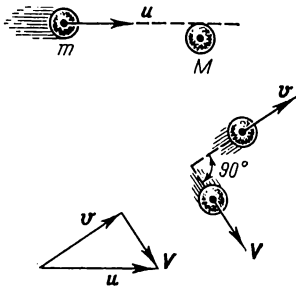


Fig. 21

A nonelastic impact is accompanied by a transformation of energy. From the example just given, it is seen that the kinetic energy may even become zero. It is not difficult to calculate the increase in the internal energy of colliding bodies in one or another case. All that we need do is perform the following subtraction:

$$\frac{m_1 + m_2}{2} V^2 - \left( \frac{m_1 v_1^2}{2} + \frac{m_2 v_2^2}{2} \right).$$

Let us now consider *ideally elastic* collisions, i.e., collisions in which the form of the bodies is completely restored. This means that no changes occur in the state of these bodies, their potential and internal energy before and after impact remain unchanged and, consequently, the kinetic energy is conserved. For two bodies colliding in this manner, two equations can be written that are based on the law of conservation of momentum and the law of conservation of kinetic energy. Let us designate the masses of the bodies by  $m$  and  $M$ . We can always make the origin of the coordinate system coincide with the position of one of the bodies. This simplifies the problem without in any way making it less general. Let us assume, therefore, that the body having mass  $M$  is at rest before impact. The above laws of conservation then yield the following two equations:

$$mu = mv + MV \quad \text{and} \quad \frac{1}{2} mu^2 = \frac{1}{2} mv^2 + \frac{1}{2} M V^2.$$

Here,  $u$  and  $v$  are the velocities of sphere  $m$  before and after impact and  $V$  is the velocity of sphere  $M$  after impact.

Let us consider several examples using these equations. First, we shall examine the case of noncentral\* collision of two spheres having equal mass (Fig. 21). The masses cancel out in both equations and we obtain

$$u = v + V \quad \text{and} \quad u^2 = v^2 + V^2.$$

From the vector equation, it is clear that the vector  $u$  closes the triangle formed by vectors  $v$  and  $V$ . The equation on the right shows that the triangle, for which

\* The impact is classified as central if the motion of the spheres before impact occurred along a straight line passing through the centres of the spheres.

is the hypotenuse, must be a right triangle. Hence, it follows that the velocities, after the collision of two particles having equal mass, must be directed at right angles to each other. This interesting conclusion is easily verified in billiards, where the directions of motion of the object ball and the cue ball form an angle of  $90^\circ$ . In other respects, the nature of the velocity change is not determined by our equations, for they do not take into consideration the deviation of the line of impact from the line passing through the centres of the spheres.

A complete description of the motion of the spheres after impact is obtained if we restrict ourselves to the case of central impact. The motion of the colliding spheres will then be along the same straight line after impact as before impact. We can dispense, therefore, with the vector notation, keeping in mind, however, that a change in the velocity's sign means that the direction of motion has changed. In this case, there is no need for making the simplifying assumption of equal masses. The equations for central collision have the form

$$mu = mv + MV \quad \text{and} \quad mu^2 = mv^2 + MV^2.$$

Rearranging terms, these equations can be written in the form

$$m(u - v) = MV \quad \text{and} \quad m(u^2 - v^2) = MV^2.$$

Dividing the latter by the former, we obtain:  $u + v = V$  or  $u = -(v - V)$ . Note that the relative velocity of motion of sphere  $m$  with respect to sphere  $M$  before impact (designated by  $u$ ) is equal in magnitude to the same relative velocity after impact.

An interesting formula is obtained when we substitute  $V = u + v$  in the formula for the law of conservation of momentum. We obtain an expression for the velocity of sphere  $m$  after impact in terms of the velocity of this sphere before impact:

$$v = \frac{m - M}{m + M} u.$$

If the masses of the spheres are equal, the velocity  $v$  reduces to zero. This phenomenon can be demonstrated very effectively with steel or ivory spheres. For such an impact, the spheres, so to speak, exchange velocities (Fig. 22). In other cases, sphere  $m$  is retarded. The closer the values of the masses of the colliding spheres, the more effective the retardation. It is not difficult to calculate that when a neutron (mass 1) rebounds from a carbon atom (mass 12) it loses  $2/13$  of its velocity and when it rebounds from a uranium atom (mass 235) it loses only  $2/236$  of its velocity.

For macroscopic bodies, the laws of elastic impact are quite valid for such materials as ivory, steel and rubber. These materials, after having been deformed, are able to reassume, to a high degree, their original form. This is illustrated by the interesting photograph shown in Fig. 23, where by means of slow-motion photography the moment of impact of a hockey ball on an obstacle is filmed. In  $1/5,000$  of a second, the ball is compressed almost one centimetre, and it takes the same amount of time for the restoring phase of the impact. In the first phase, the kinetic energy of the impact is transformed into potential energy of elastic compression. In the second phase, the potential energy is transformed back into kinetic energy. For an ideal impact, this reverse process should completely restore the value of kinetic energy expended during the first phase of the impact.

Our formulas are not applicable for the important case of elastic impact of a sphere on a wall (Fig. 24). Since the kinetic energy must be conserved, the velocity of the sphere cannot change in magnitude. As regards the direction of the

sphere's motion after impact, it should form the same angle with the normal ( $90^\circ - \alpha$ ) as before impact. Thus, in the case of impact on a smooth wall, the tangential component of the velocity remains unchanged, since no tangential adhesion forces are exerted by the wall. As can be seen from the figure, the increment of the momentum is numerically equal to  $2mv \sin \alpha$  and is directed along the normal to the

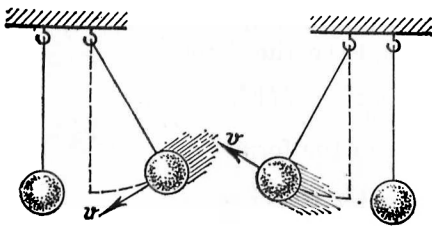


Fig. 22

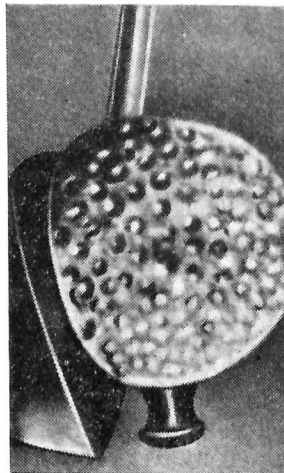


Fig. 23

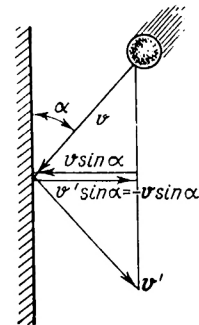


Fig. 24

wall. According to the fundamental law of mechanics, at the instant of impact, the force exerted on the sphere by the wall has the same direction as that of the vector of momentum change. The angle of incidence of the sphere is, therefore, equal to its angle of reflection.

Let us consider an inelastic impact, using as our example a ballistic pendulum (a device for measuring the velocity of a bullet). A bob containing sand, mass  $M$ , hangs from a line. The bullet is fired into the bob and becomes imbedded in the sand. The momentum of the bullet before impact is  $mu$ , and the momentum of the system after impact is  $(M + m)v$ . Hence,

$$v = \frac{m}{m + M} u.$$

Acquiring the kinetic energy  $\frac{Mv^2}{2}$ , the bob expends it in rising to the height  $h$ , determined by the following condition:

$$Mgh = \frac{Mv^2}{2}; \quad \text{i.e.,} \quad h = \frac{u^2}{2g} \left( \frac{m}{M} \right)^2 \quad (m \ll M).$$

If  $M = 10$  kg,  $m = 10$  g and  $u = 900$  m/sec, then  $h = 4$  cm.

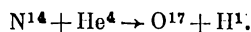
If we had not used the law of conservation of momentum in determining  $h$ , but had assumed instead that the total kinetic energy of the bullet had been transformed into potential energy of the pendulum we would have obtained the value  $h = 40$  metres (!). This means that, in our example, 3920 J of mechanical energy, or 99.9% of the total supply, has "disappeared", i.e., gone to heat the system. Since absolutely elastic bodies do not exist, mechanical energy is not conserved for "elastic" impacts as well, for part of it is transformed into energy of thermal molecular motion and is dissipated. We shall return to this example in Chapter XI (p. 135).

Using an example involving collision we shall now illustrate the merit of a coordinate system bound to the centre of mass.

Assume that a sphere of mass  $m$ , at rest in a laboratory coordinate system, is hit by a similar sphere with velocity  $v$ . If the impact is inelastic, some portion of  $\frac{mv^2}{2}$ , the kinetic energy of the system, is transformed into heat. In other coordinate systems, the kinetic energy of this pair of spheres is expressed by other quantities. As regards the heat released, it will be the same for the given pair of spheres and is simply determined by the velocity of their *relative* motion. Therefore, instead of resorting to the law of conservation of momentum to try to determine the portion of the kinetic energy that is transformed into heat, calculated for the laboratory coordinate system, it is sufficient to calculate the kinetic energy for a coordinate system bound to the centre of mass. Since in such a coordinate system the total momentum of bodies is equal to zero, after an inelastic collision the spheres come to a standstill: all the kinetic energy is transformed into heat. The kinetic energy will have a minimal value for a system bound to the centre of mass.

In a coordinate system bound to the centre of mass, the spheres move toward each other with velocities  $\frac{1}{2}v$ . The kinetic energy of each sphere is thus equal to  $\frac{1}{4}mv^2$  and the total energy of the system is  $\frac{1}{2}mv^2$ . This is the amount of heat released during an inelastic collision. Irrespective of the type of impact, the heat (or other form of energy) released due to the kinetic energy of the bodies cannot exceed the amount of kinetic energy calculated for a system bound to the centre of mass. And conversely, in order to release a given amount of heat, it is necessary to have the equivalent amount of kinetic energy calculated for a centre of mass system.

*Example.* A nuclear reaction in which  $\alpha$ -particles bombard nitrogen  $N^{14}$  takes place in accordance with the following equation:



The energy absorbed in this process amounts to 1.13 MeV. How much kinetic energy in a laboratory system must an  $\alpha$ -particle possess in order for the reaction to proceed? At first glance, it seems that 1.13 MeV is sufficient for this purpose. But we already know that this is not the case. In a centre of mass coordinate system, 1.13 MeV is required, but in a laboratory coordinate system, more energy is needed.

Thus, the velocity of the centre of mass is  $v_c = \frac{m_1 v_1 + m_2 v_2}{m_1 + m_2}$ , where  $m_1 v_1$  is the momentum of the first particle and  $m_2 v_2$  is the momentum of the second. The velocity of the first particle in a centre of mass coordinate system is  $v'_1 = v_1 - v_c = \frac{m_2}{m_1 + m_2} (v_1 - v_2)$ . For the second particle, we may write:  $v'_2 = v_2 - v_c = \frac{m_1}{m_1 + m_2} (v_2 - v_1)$ . Hence, the kinetic energy of the system ( $\alpha$ ,  $N^{14}$ ) in a centre of mass coordinate system is  $K_{cm} = \frac{1}{2} \mu (v_1 - v_2)^2$ , where  $\mu = \frac{m_1 m_2}{m_1 + m_2}$  is the so-called reduced mass of both particles. We shall consider the nuclei of  $N^{14}$  fixed ( $v_2 = 0$ ). This assumption is justified, since we can always neglect the slow thermal motion of the target nuclei as compared with the large velocity of the bombarding particles. The kinetic energy in the laboratory coordinate system is then  $K_{lab} = \frac{1}{2} m_1 v_1^2$  and therefore

$$K_{lab} = K_{cm} \frac{m_1 + m_2}{m_2}.$$

The reaction proceeds if  $K_{cm} = 1.13$  MeV. Since  $m_1 = 4$  and  $m_2 = 14$ , we obtain

$$K_{lab} = 1.13 \times \frac{18}{14} = 1.45 \text{ MeV.}$$

## Sec. 17. RECOIL

The law of conservation of momentum helps one to easily understand the fundamentals of recoil in gunfire, reaction propulsion, and other similar phenomena.

We shall consider, in the first place, recoil taking place in a frame of reference where the bodies are at rest at the initial moment. In the case of gunfire, this assumption is in complete accord with the prevailing conditions. If at the initial moment a system consisting of two or more bodies is at rest, the total momentum of the system is equal to zero. Irrespective of the future course of events, the total momentum continues to be equal to zero. Thus, if at some instant an explosion takes place, causing the system to be divided into parts having masses  $m_1, m_2, m_3, \dots$ , which fly asunder with velocities  $v_1, v_2, v_3, \dots$ , the total momentum  $m_1v_1 + m_2v_2 + m_3v_3 + \dots$  of the scattered bodies must be, as before, equal to zero.

In the case of gunfire (where the system divides into two parts), the condition that the momentum of this system of two bodies be equal to zero has the form  $mv + MV = 0$ . Here, the lower-case letters refer to one body, say the missile, and the capital letters to the other—the gun. The division of the system into two parts can only take place along a straight line. We can, therefore, dispense with the vector notation and write the condition in the form  $mv = -MV$ . The velocities of the gun and the missile are inversely proportional to their masses. Thus, the greater the mass of the missile with respect to the mass of the gun, the greater the observed recoil.

The phenomenon of “continuous recoil”, occurring in reaction propulsion, is of exceptional interest. It is the subject of a distinctive branch of mechanics that may be called the mechanics of variable mass. This phenomenon does not only occur in jet planes. Indeed, we can point to a number of commonplace occurrences involving such motion. As examples, it is sufficient to mention the case of an uncoiling roll of paper or the fall of droplets continuously condensing in the atmosphere (see the example at the end of this section). The fundamentals of the mechanics of variable mass were developed at the end of the nineteenth century by Prof. I. V. Meshchersky. Since we cannot describe his work here, we shall restrict ourselves to the consideration of a single problem in this field—a problem related to the possible velocity of motion of a rocket.

A rocket moves with a velocity  $v$  and at some instant ejects a certain amount of combustible gas having mass  $dM$ . The mass of the rocket, naturally, decreases by this amount. If the velocity of the ejected gas is designated by  $u$  (this velocity is not given with respect to the rocket, but with respect to the inertial coordinate system in which the velocity of the rocket motion is described), the momentum of the matter escaping from the rocket will be equal to  $u dM$ . The rocket decreases its mass and increases its velocity by the amount  $dv$ . The momentum of the rocket after ejecting the fuel is equal to  $(M - dM)(v + dv)$ . In accordance with the law of conservation of momentum, we can equate the momentum  $Mv$  of the rocket before discarding a portion of the fuel and the momentum of the system after that quantity of gas has been ejected. The latter is equal to the difference between the momentum of the rocket and the momentum of the portion of fuel. Thus,

$$Mv = (M - dM)(v + dv) - u dM.$$

Whence, excluding second-order infinitesimals,

$$dv = (u + v) \frac{dM}{M}.$$

But  $u + v$  is the relative velocity of the outflowing combustible gas (with respect to the rocket). Designating this velocity by  $c$ , we arrive at the following equation for the increment in the rocket's velocity:  $dv = -c \frac{dM}{M}$ . The minus sign is used to show that the velocity increases when the mass decreases. It can be seen that the increase in velocity is equal to the fraction of the lost mass multiplied by the relative velocity of the ejected fuel.

Taking the velocity of the outflowing gas with respect to the rocket to be a constant value, the above equation can be easily integrated. If the mass of the rocket was  $M_0$  when the velocity of the rocket was  $v_0$ , and became equal to  $M$  when the velocity of the rocket changed to  $v$ , integration yields

$$\int_{v_0}^v dv = -c \int_{M_0}^M \frac{dM}{M},$$

i.e.,

$$v - v_0 = c \ln \frac{M_0}{M}.$$

The latter formula was initially obtained by K. E. Tsiolkovsky, the first to design a rocket and do research in the theory of interplanetary travel.

Going over to common logarithms and introducing the designation  $m = M_0 - M$  for the difference in the mass of the rocket, i.e., for the mass of the ejected fuel, we obtain Tsiolkovsky's formula in the form

$$v = c \times 2.3 \times \log \left( 1 + \frac{m}{M} \right)$$

(the initial velocity  $v_0$  is assumed to be equal to zero).

In modern rockets, the velocity of gas outflow is probably not less than 2,000 m/sec. Using this velocity in the formula, the following table of values is obtained:

$\frac{m}{M}$	0.25	1.0	4.0	0.0	32.3	54	999
$v$ (m/sec)	446	1,386	3,218	4,817	7,013	8,000	13,815

As can be seen from this table, the rocket velocity increases much slower with respect to the amount of ejected fuel than one would like. To give the rocket a large velocity, a tremendous amount of fuel, relative to the initial mass of the rocket, must be ejected. Thus, if a velocity of 7 km/sec is imparted, less than  $\frac{1}{30}$  of the initial rocket mass will remain.

A velocity of about 11 km/sec must be imparted to a rocket if it is to escape from the Earth's gravitational pull. This figure is obtained in the following simple manner. To escape from the Earth, a rocket must possess sufficient kinetic energy to perform the work of moving a body from the Earth's surface to infinity. But this work against the force of gravity is equal to the difference between the rocket's potential energy at the Earth's surface and at infinity. Since at infinity the potential energy is equal to zero, the condition for escape from the Earth has the following simple form:

$$\frac{mv^2}{2} = \gamma \frac{mM}{R},$$

where  $M$  and  $R$  are the Earth's mass and radius, respectively. Multiplying the numerator and the denominator of the right-hand member of the equation by  $R$ ,

then substituting  $\gamma \frac{M}{R^2}$  by  $g$ , the acceleration of the gravitational force at the Earth's surface, and cancelling the rocket's mass, we obtain the condition for escape from the Earth:  $v = \sqrt{2gR}$ , which yields a figure of about 11 km/sec.

If we assume that the velocity of the gas outflow is 2,000 m/sec,\* the ratio  $\frac{m}{M}$  can be obtained from Tsiolkovsky's formula. It is equal to 244. For the rocket to escape from the Earth, its design must be such that only  $\frac{1}{245}$  of the rocket's mass before take-off will remain in its interplanetary flight. If we were to succeed in increasing this velocity by a factor of three, i.e., increasing this velocity to 6 km/sec, the ratio  $\frac{m}{M}$  would fall to 5.3. But this is, evidently, unreal at present judging, say, by the press information published in December, 1968 concerning the flight of "Apollo-8" round the Moon: "a descending compartment (modulus) weighing 5.3 tons will return into the Earth's atmosphere—all what will remain from a 3100-ton spaceship".

It is easier to put an Earth satellite into orbit because a smaller initial velocity is required. If we assume that the acceleration of the gravitational force at the heights at which we desire the satellite to orbit is approximately the same as at the Earth's surface, then the law of mechanics for artificial planets will have the form  $mg = ma$ , and since the satellite moves in a circle, the centripetal force is  $a = \frac{v^2}{R}$ . Thus, the velocity of the rotating satellite is  $v = \sqrt{gR}$ , i.e., 8 km/sec. When such a velocity is imparted to a rocket, it is transformed into an Earth satellite. From the above table, calculated for the velocity of gas outflow of 2,000 m/s, we see that the value of  $\frac{m}{M}$  required for imparting a velocity of 8 km/sec to a rocket is 54.

*Example of motion of a body with variable mass.* Consider a water droplet falling in an atmosphere saturated with water vapour. At the instant of time  $t$ , the droplet has a mass  $m$  and a radius  $r$ . During the time  $dt$ , the volume of the droplet, and hence the mass (for a density equal to 1), increases by  $4\pi r^2 dr$ . Thus, the rate of increase in mass is  $\frac{dm}{dt} = 4\pi r^2 \frac{dr}{dt}$ . At the same time, it is clear from physical considerations that  $\frac{dm}{dt}$ , the rate of condensation of the water vapour, must be proportional to  $4\pi r^2$ , the condensation surface. Hence,  $\frac{dr}{dt} = \text{const}$  and  $r = kt$ , where  $k$  is some constant of proportionality.

Let us derive the equation of motion of this droplet in the Earth's gravitational field. We are interested in the change of momentum  $d(mv)$ , which according to the fundamental law of mechanics is equal to  $Fdt$ , where  $F = mg$ . Thus:

$$F = \frac{d}{dt}(mv), \quad \text{i.e.,} \quad mg = m \frac{dv}{dt} + v \frac{dm}{dt}.$$

Substituting the expressions for  $m$  and  $r$ , we obtain

$$\frac{dv}{dt} = g - \frac{3v}{t}.$$

By integrating this equation, we arrive at the following result:  $v = \frac{g}{4}t$ , i.e., the droplet falls with the constant acceleration  $\frac{1}{4}g = 2.45 \text{ m/sec}^2$ . The resistance of the air was not taken into consideration.

---

\* It was reported by the Soviet sources that a liquid engine is capable of giving outlet velocities up to 4,500 m/s.



# Rotation of a Rigid Body

## Sec. 18. KINETIC ENERGY OF ROTATION

In this chapter, we shall be concerned with “perfectly rigid” bodies. This means that we may neglect any deformation occurring during the motion of such a body, and assume that the distances between the particles of the body remain unchanged.

Let us consider a rigid body rotating about a fixed axis passing through it (Fig. 25). We can conceive of the body as consisting of small volumes of masses  $\Delta m_1, \Delta m_2, \dots$  at distances  $r_1, r_2, \dots$  from the axis of rotation. Corresponding to the various values of the distances are the various velocities of motion  $v_1, v_2, \dots$ . We are interested in the kinetic energy of rotation of the entire rigid body, which is composed of the kinetic energies of the individual particles  $\Delta m_1, \Delta m_2, \dots$ , i.e.,

$$K_{rot} = \frac{\Delta m_1 v_1^2}{2} + \frac{\Delta m_2 v_2^2}{2} + \frac{\Delta m_3 v_3^2}{2} + \dots$$

The velocity of angular motion of any point of the body can be easily expressed in terms of  $\omega$ , the angular velocity of the rotating body. If the body turns through an angle  $d\varphi$  in the time  $dt$ , the derivative  $\frac{d\varphi}{dt}$  is called the *angular velocity*:

$$\omega = \frac{d\varphi}{dt}.$$

For the case of uniform motion, the above formula is transformed into a relation already known to the reader, namely,  $\omega = \frac{2\pi}{T}$ . The quantity  $\omega$  is usually measured in radians per second. If the body performs 1 revolution per second, its angular velocity is equal to  $2\pi$  rads/sec.

Different points of a rotating rigid body have different velocities  $v$  (called *linear velocities*), but the angular velocity  $\omega$  is the same for each point. In turning through an angle  $d\varphi$ , a point describes an arc  $ds = r d\varphi$ . Dividing both members of this equation by the time of motion  $dt$ , we obtain a relationship between linear and angular velocity:

$$v = \omega r.$$

Thus, the formula already known for uniform motion is valid in the general case.

Using this relation, the expression for  $K_{rot}$  may be written in the following form:

$$K_{rot} = \frac{\omega^2}{2} (r_1^2 \Delta m_1 + r_2^2 \Delta m_2 + \dots).$$

The quantity in brackets does not depend on the velocity of motion, but is a measure of the inertial properties of a body executing rotational motion. The greater the value of the expression in brackets, the greater the energy that must be expended to achieve a given velocity. Therefore, the quantity

$$I = r_1^2 \Delta m_1 + r_2^2 \Delta m_2 + \dots$$

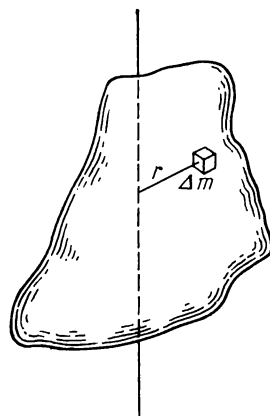


Fig. 25

is called *the moment of inertia* of the body, and the term  $r^2 \dot{m}$  is *the moment of inertia of a particle*. The quantity  $I$  may be expressed more briefly as follows:

$$I = \int r^2 dm,$$

where the integration (summation) encompasses all the particles of the body.

The formula for a body's kinetic energy of rotation acquires the form

$$K_{rot} = \frac{I\omega^2}{2}.$$

This formula is valid for a body rotating about a fixed axis. For a rolling body (a ball, wheel, etc.), the energy of motion will consist of the energy of rotational and translational motion. Thus, if a rolling body has a mass  $M$ , moment of inertia  $I$ , translational velocity  $v$  and rotational velocity  $\omega$ , the kinetic energy is

$$K_{roll} = \frac{Mv^2}{2} + \frac{I\omega^2}{2}.$$

Moreover, it turns out that this formula is valid for any arbitrary motion of a rigid body. In theoretical mechanics, it is shown that any arbitrary motion can always be resolved into translational and rotational motion. The rotation, in this case, must be considered with respect to an axis passing through the centre of mass.

#### Sec. 19. MOMENT OF INERTIA

If we carefully examine the formula for moment of inertia, we see that the value of  $I$  depends on the nature of the distribution of the mass with respect to the axis of rotation. The particles that are far from the axis of rotation contribute considerably more to the total value than those that are close to it.

Let us calculate the moment of inertia of a flat disk, of radius  $r$ , relative to the axis perpendicular to the plane of the disk and passing through its centre (Fig. 26).

The mass of an annular element of radius  $x$  is  $dm = \rho \times 2\pi x dx$ , where  $\rho$  is the density of the disk's material. This ring has a moment of inertia  $dI_1 = dm x^2$ , and the moment of inertia of the entire disk is:

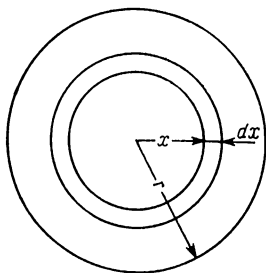


Fig. 26

$$I_1 = \int_0^r dI_1 = \int_0^r \rho \times 2\pi \times x^3 dx = 2\pi\rho \frac{r^4}{4} = \frac{\pi\rho r^4}{2}.$$

It is evident that with respect to the same axis the moment of inertia of a ring is  $I_2 = mr^2$ , i.e.,  $I_2 = 2I_1$ , when the entire mass is concentrated at the outer circumference.

The moment of inertia of a body will vary in accordance with the location of the axis of rotation. If a thin rod rotates about its long axis, the moment of inertia will be very small, for all the particles lie very close to the axis of rotation and, therefore, all the quantities  $r_1^2, r_2^2, \dots$  entering in the formula for  $I$  will have very small values. The moment of inertia will be much larger if the rod is rotated about a line perpendicular to its axis.

The moment of inertia depends on the orientation of the axis and the location of the point through which it passes. If no specific stipulation to the contrary is made, it is assumed that the axis of rotation passes through the body's centre of mass.

If the axis of rotation is displaced relative to the centre of mass by the amount  $a$  (Fig. 27),  $I$ , the new moment of inertia, will differ from  $I_0$ , the moment of inertia with respect to the parallel axis passing through the centre of mass.

In view of what was stated at the end of the previous article, we can express the kinetic energy of a body rotating about the displaced axis as the sum

$$K = \frac{Mv^2}{2} + I_0 \frac{\omega^2}{2},$$

where  $v$  is the velocity of motion of the centre of mass and is equal to  $a\omega$ . Thus,

$$K = \frac{Ma^2\omega^2}{2} + \frac{I_0\omega^2}{2} = \frac{\omega^2}{2} (I_0 + Ma^2).$$

Hence, the moment of inertia  $I$  with respect to a parallel axis, displaced by a distance  $a$  from the centre of mass, can be expressed as follows:

$$I = I_0 + Ma^2.$$

It follows that the moment of inertia with respect to an axis passing through the centre of mass is always the smallest possible for a given orientation. Depending on its symmetry, a body will have one, two or three moments of inertia with respect to the main axes passing through the centre of mass.

Thus, a disk is characterized by two axes passing through its centre—one lying in the plane of the disk and the other perpendicular to the disk. The moments of inertia are then  $\frac{mr^2}{4}$  and  $\frac{mr^2}{2}$ , respectively (it is assumed, naturally, that the distribution of mass throughout the disk is uniform). For a ring, the moment of inertia about similar axes is  $\frac{mr^2}{2}$  and  $mr^2$ , respectively.

For all solids of revolution, it is sufficient to know the moments of inertia with respect to two axes. In the case of a body of arbitrary form, to completely describe the inertial properties of the body during rotation, it suffices to know three moments of inertia with respect to axes passing through the centre of mass, namely,  $I_{\max}$ —the largest moment of inertia,  $I_{\min}$ —the smallest, and  $I_{\text{mean}}$ —the moment of inertia with respect to an axis perpendicular to the first two.

The only body for which the moment of inertia about all the axes is the same is a sphere. For a sphere,  $I = \frac{2}{5} mr^2$ .

The above formulas for moment of inertia are calculated from the relation:

$$I = \int r^2 dm.$$

To use this formula, it is generally necessary to be able to operate with multiple integrals. Examples of such calculations are given in courses on theoretical mechanics.

As we shall see below, physicists are sometimes interested in the values of moments of inertia for molecules. Since the mass of atoms is concentrated in nuclei whose dimensions are very small, the calculation of the moments of inertia can be accomplished without difficulty, for the atoms may be considered as point masses.

For a diatomic molecule, the moment of inertia with respect to the axis passing through the atoms is equal to zero. For the axis perpendicular to the line joining

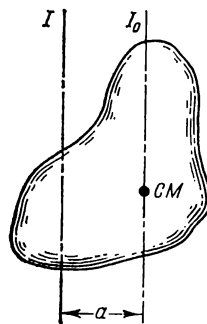


Fig. 27

the atoms we obtain

$$I = m_A r_A^2 + m_B r_B^2,$$

where  $r_A$  and  $r_B$  are the distances of atoms  $A$  and  $B$  of a diatomic molecule to the centre of mass. If  $l$  is the distance between atoms,  $r_A + r_B = l$  and  $\frac{r_A}{r_B} = \frac{m_B}{m_A}$ . Therefore,

$$I = \frac{m_A m_B}{m_A + m_B} l^2.$$

The moments of inertia of more complex molecules may also be calculated as the sum of the moments of inertia of the atoms considered as point masses.

*Examples.* 1. The flywheel of a ship's engine has a mass of about 1 ton, a diameter of 2 metres and, therefore, a moment of inertia  $I \sim 1,000 \text{ kg m}^2$ . Making 300 rpm, the flywheel possesses a kinetic energy of rotation

$$K = \frac{I\omega^2}{2} \approx 500,000 \text{ J} \approx 50,000 \text{ kgf-m.}$$

2. The moment of inertia of the Earth is about  $10^{45} \text{ g cm}^2 = 10^{38} \text{ kg m}^2$ . The kinetic energy of rotation of the Earth about its axis is  $2.5 \times 10^{29} \text{ J}$ .

3. In a molecule of hydrogen  $\text{H}_2$ , the distance  $l = 0.753 \times 10^{-8} \text{ cm}$ , the mass of the hydrogen atom  $m_{\text{H}} = 1.6598 \times 10^{-24} \text{ g}$  and, therefore, the moment of inertia of the molecule with respect to the axis perpendicular to  $l$  is

$$I = \frac{m_{\text{H}} l^2}{2} = 0.46 \times 10^{-40} \text{ g cm}^2.$$

## Sec. 20. ROTATIONAL WORK AND THE FUNDAMENTAL EQUATION OF ROTATION

If a body fixed on a shaft is made to rotate by a force  $F$  or, on the other hand, if a rotating body is braked by the force  $F$ , the kinetic energy of rotation increases or decreases by the magnitude of the expended work. Just as in the case of translational motion, this work depends on the effective forces and the displacement produced thereby. However, the displacement is now angular, and the expression that we know for the displacement of a particle by a certain distance is not applicable here.

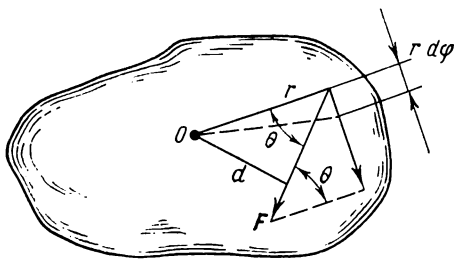


Fig. 28

To find the formula that we are interested in, let us refer to Fig. 28. The force  $F$  is applied at a point located at a distance  $r$  from the axis of rotation. The angle between the direction of the force and the radius vector is designated by  $\theta$ . Since the body is perfectly rigid, the work of this force (even though applied at one point)

is equal to the work expended in rotating the entire body. In rotating the body through an angle  $d\varphi$ , the point of application traverses the path  $r d\varphi$  and the work  $dA$ , equal to the product of the projected force along the direction of displacement and the magnitude of the displacement, is then

$$dA = Fr \sin \theta d\varphi.$$

$Fr \sin \theta$  is known as the *moment of force* or *torque*:  $M = Fr \sin \theta$ . From the diagram, it is seen that  $r \sin \theta = d$ , where  $d$  is the shortest distance between the line

of action of the force and the axis of rotation. Hence,

$$M = Fd,$$

i.e., the torque is equal to the product of the force and the lever arm.

The formula for work that we have sought is

$$dA = M d\varphi.$$

The work of rotating a body is equal to the product of the effective torque and the angle of rotation.

Strictly speaking, the formula is only valid for an infinitely small angle  $d\varphi$ . However, we may use it in any case if we understand  $M$  to mean the average value of the torque for the time of rotation. Then,

$$\Delta A = M_{av} \Delta\varphi.$$

The work of rotation goes to increase the kinetic energy of rotation. Hence, the following equation must hold:

$$M d\varphi = d \left( \frac{I\omega^2}{2} \right).$$

If the moment of inertia is constant for the time of motion, then

$$M d\varphi = I\omega d\omega$$

or, since  $\omega = \frac{d\varphi}{dt}$ ,

$$M = I \frac{d\omega}{dt}.$$

This is the fundamental equation of motion for a rotating body. The torque acting on a body is equal to the product of the moment of inertia and the angular acceleration  $\frac{d\omega}{dt}$ .

*Examples.* 1. The torque on a wheel of a locomotive developing a traction of about  $10^5$  N is about 3,000 N·m.

A man riding a bicycle produces a torque of about 100 N·m on the pedals.

2. By means of an example, we shall show the connection between the expression for the kinetic energy of a moving rigid body (see p. 55) and the fundamental law of mechanics.

Let us consider a spool of mass  $m$  and radius  $r$ , possessing a moment of inertia  $I$  with respect to its axis and wound with weightless thread (Fig. 29). The free end of the thread is fastened at a certain height above the Earth's surface. The spool is allowed to fall under the action of its own weight  $mg$ . Hence, the equations of motion for the spool are:

$$mg - T = m \frac{dv}{dt}$$

and

$$Tr = I \frac{d\omega}{dt},$$

where  $T$  is the tension of the thread and  $\omega$  is the angular velocity of rotation of the spool. Eliminating  $T$ , we obtain for the acceleration:

$$a = \frac{dv}{dt} = \frac{g}{1 + \frac{I}{mr^2}}.$$

If time is counted from the moment the spool begins to fall, then in  $t$  seconds the spool will fall a distance  $h = \frac{v^2}{2a}$ . It is evident that the total kinetic energy of the spool at that instant is equal

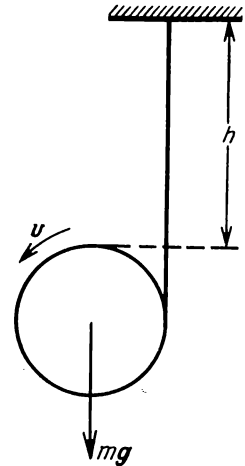


Fig. 29

to the change in the potential energy of the spool:

$$K = mgh = mg \frac{v^2}{2a}.$$

Substituting the expression for  $a$ , we obtain

$$K = \frac{mv^2}{2} + \frac{I\omega^2}{2}.$$

## Sec. 21. ANGULAR MOMENTUM

The similarity between the formulas of motion of a particle and the derived laws for the rotation of a rigid body is immediately evident. Thus, compare the following formulas:

<i>Particle</i>	<i>Rotating body</i>
$F = m \frac{dv}{dt},$	$M = I \frac{d\omega}{dt},$
$K = \frac{mv^2}{2},$	$K = \frac{I\omega^2}{2}.$

Clearly, the physical concepts are also analogous. While in the mechanics of particles the acceleration is determined by the force, in rotational motion the angular acceleration is determined by the moment of force, i.e., the torque. The role of mass is played by the moment of inertia, which in rotation is the measure of a body's inertia (the mass alone is insufficient here for this purpose). This similarity encourages us to go a step further and assume that analogous physical quantities are related by analogous relations.

In the previous chapter, it was established that the momentum  $p = mv$  is a physical quantity satisfying the law of conservation in a closed system. The quantity analogous to  $p$  is the *moment of momentum (angular momentum)*:

$$N = I\omega.$$

It can be rigorously proved that angular momentum satisfies the law of conservation, i.e., in a closed system, the total angular momentum of the bodies belonging to this system does not change. An increase in the angular momentum of one of the bodies is compensated for by an equivalent decrease in the others.

The relation

$$I_1\omega_1 + I_2\omega_2 + I_3\omega_3 + \dots = \text{const}$$

has many interesting applications that are in many ways analogous to the problems studied in the previous chapter.

The law of conservation of momentum when applied to a single body has the form  $mv = \text{const}$  and is, therefore, identical with the law of inertia. Even in this simple case, the law of conservation of angular momentum leads to an interesting result. A single body, in the absence of interaction with its medium, must satisfy the condition

$$I\omega = \text{const}.$$

However, the moment of inertia of a body may change during motion. It is, therefore, evident that an increase in  $I$  must be accompanied by a decrease in  $\omega$ , and vice versa.

One can cite numerous examples, and this phenomenon can be strikingly demonstrated by means of a swivel stool. Holding a pair of dumb-bells in your hands,

be seated on such a stool (Fig. 30). Place your arms outstretched in a horizontal position and have someone give you a small rotatory push. Motion takes place for the particular moment of inertia  $I$  at an angular velocity  $\omega$ . Now fold your arms on your chest. As a result the moment of inertia drops sharply to  $I'$ . Since the

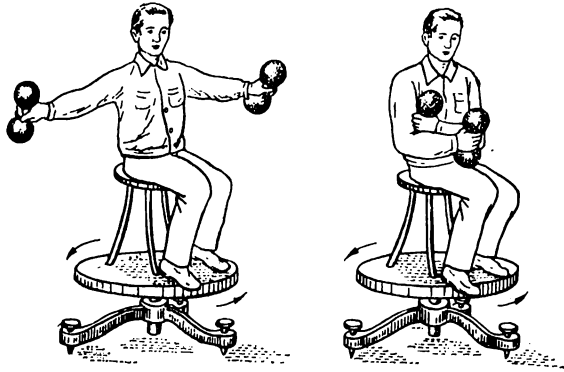


Fig. 30

product  $I\omega$  remains unchanged,  $I\omega = I'\omega'$ . Thus, changing the position of one's arms leads to a considerable increase in the velocity of rotation. The process may be repeated—stretching one's arms out leads to retarded motion and folding them produces accelerated motion.

Decreasing the moment of inertia as a method of increasing the velocity of rotation is quite familiar to gymnasts and dancers. It is used in all kinds of jumps,



Fig. 31

tumbles and spins. Thus, a ballet dancer in a position of large moment of inertia will impart velocity by changing her posture to a position of small moment of inertia (Fig. 31).

Rotational recoil is usually demonstrated by means of the afore-mentioned swivel stool and a wheel fixed on a long axle (Fig. 32). While standing on the stool

and holding the wheel above one's head, the wheel is twirled by means of a sudden movement. As a result, the stool rotates in the direction opposite to that of the wheel. This is precisely what is meant by recoil.  $I_1\omega_1$ , the angular momentum of the wheel, is balanced by  $I_2\omega_2$ , the angular momentum of the stool with the person standing on it (the angular momenta have opposite signs). This is due to the fact that in the initial state both the stool and the wheel did not rotate, and the total angular momentum was equal to zero.



Fig. 32

An inelastic impact was defined above as an encounter between two bodies as a result of which the bodies move together. Something analogous may be demonstrated in the case of rotation using the equipment just described. The wheel is made to rotate and is then transferred to a person standing on the stool. Thus, the initial state is the following: the stool and the person standing on it are at rest, while the bicycle wheel is rotating with a momentum  $I_1\omega_1$ . Now, the person on the stool takes hold of the wheel. The angular momentum  $I_1\omega_1$  cannot disappear, but it now belongs to the entire system. Naturally, the person on the stool and the wheel rotate together in the same direction as the wheel was rotating. Clearly,  $I_1\omega_1 = (I_1 + I_2)\omega$ . If before "unification" the person rotated with a velocity  $\omega_2$ , the angular momentum to be conserved is  $I_1\omega_1 + I_2\omega_2$ . Therefore,

$$I_1\omega_1 + I_2\omega_2 = (I_1 + I_2)\omega \quad \text{or} \quad \omega = \frac{I_1\omega_1 + I_2\omega_2}{I_1 + I_2}.$$

This is very similar in form and in content to the expression for inelastic impact.

*Examples.* 1. The flywheel of a ship's engine has a moment of inertia of 1,000 kg m<sup>2</sup> and at 300 rpm its angular momentum is  $\sim 30,000 \frac{\text{kg m}^2}{\text{sec}}$ .

2. A billiard ball whose radius is 2.5 cm has a moment of inertia  $I = 250 \text{ g cm}^2$  and moves with a velocity of 5 m/sec without skimming the table. Its angular momentum is then  $\sim 50,000 \text{ g cm}^2/\text{sec} = 5 \times 10^{-3} \text{ kg m}^2/\text{sec}$ .

3. The angular momentum of the Earth in rotating about its axis is  $\sim 10^{34} \text{ kg m}^2/\text{sec}$ .

## Sec. 22. FREE AXES OF ROTATION

Let us assume that a body has received angular momentum about some axis to which the body is fastened. Further, let us assume that the fastening is then removed. While the angular momentum must be conserved (naturally, neglecting friction), the orientation of the body in space may change. If this occurs, and as a result there is a change in the moment of inertia, it will be compensated for by a corresponding change in the angular velocity.

However, in a number of cases the nature of the rotation does not change. Stable rotation takes place about the original axis, just as if the axis of rotation were fixed as before. Theory and experiments show that there are two axes passing through the centre of mass that may be permanent, free axes of rotation, namely, the axis of maximum moment of inertia and the axis of minimum moment of inertia.

If the fixed axis of rotation passes through the centre of mass (Fig. 33), but is inclined to the axes of symmetry and, therefore to the afore-mentioned orienta-



tions, then after the fastening is removed, the body begins to change its orientation with respect to the axis of rotation. It can be seen from the figure that the reason for the change of orientation is the fact that the centrifugal forces form a couple of forces. The body will continue changing its orientation until the axis of rotation becomes a free axis.

It can be shown in a number of ways that a freely rotating body will keep changing its axis of rotation until the rotation occurs about a free axis. Tying bodies of various shapes to one end of a string, and attaching the other end to the shaft of a rapidly rotating motor, we can transmit rotary motion to a body without having a fixed axis of rotation. In Fig. 34, the successive orientations of a rotating hoop,

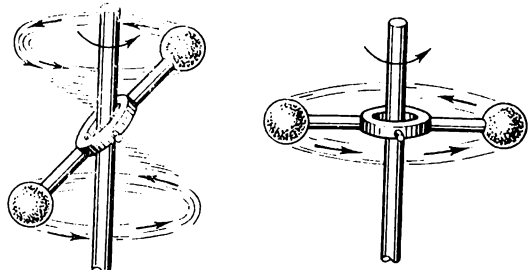


Fig. 33

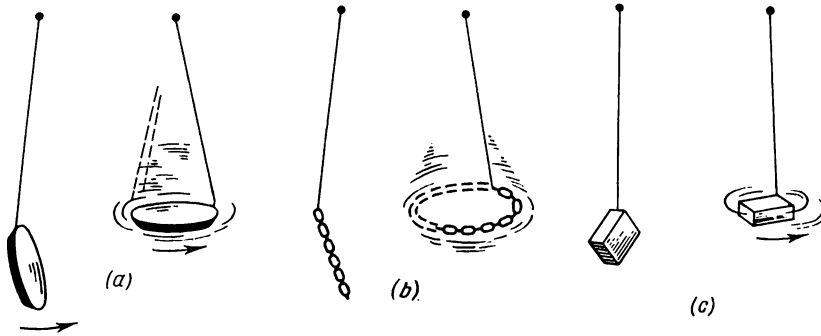


Fig. 34

chain and match box are shown. The match box will begin rotating about its shortest or about its longest edge. Theory shows that rotation about the axis having a mean moment of inertia will not be stable even if this axis is an axis of symmetry.

In constructing one of the first turbines, attempts to fix the position of the shaft with sufficient accuracy to eliminate the couple of centrifugal forces acting on the bearings at a velocity of 30,000 rev/min were unsuccessful. At such high velocities, these forces are intolerably great. The problem was solved by using a flexible shaft for the turbine wheel. The rotation took place about the free axis and the flexible shaft adapted itself to this axis.

Let us consider this phenomenon in somewhat more detail. We shall designate the shift in the centre of gravity of the turbine wheel due to the wheel's asymmetry by  $a$  and the amount by which the shaft sags under the action of the centrifugal force by  $\Delta$ . The shaft sags in the direction of the asymmetry. Hence, the expression for the centrifugal force may be written in the form  $4\pi^2 n^2 M (a + \Delta)$ . This force is balanced by the elastic force  $k\Delta$ , where  $k$  is the stiffness of the shaft. Thus,

$$\Delta = a \frac{1}{\frac{k}{4\pi^2 n^2 M} - 1}.$$

The formula shows that when the number of revolutions per minute  $n$  is large the shaft's sag  $\Delta$  does not increase, but tends to become equal to minus  $a$ , the measure of the wheel's asymmetry. This means that when the angular velocity of the turbine increases the total displacement of the wheel with the shaft from the axis of rotation tends to become equal to zero. Herein lies the

adaptability of the flexible shaft: It can bend, without breaking, by the amount required to eliminate the centrifugal force.

From the above formula, it follows that the condition  $k/4\pi^2 n^2 M = 1$  is critical, for the relation shows that the shaft's sag becomes infinitely large. This is the instant of resonance which must be rapidly passed in running the turbine (the external frequency  $n = \frac{1}{2\pi} \sqrt{\frac{k}{M}}$ , i.e.,  $n$  coincides with the natural frequency of a turbine wheel of mass  $M$  placed on a shaft of stiffness  $k$ ; see Chapter V).

### Sec. 23. THE GYROSCOPE

The term gyroscope usually denotes a device that can rotate about any orientation of its axis. If a gyroscope is rotated and then not interfered with, its axis of rotation will remain unchanged as long as no forces act on it ( $I\omega$ , in this case, should not change).

The action of a force on a gyroscope's axis of rotation manifests itself in a somewhat surprising manner. This may be demonstrated using a gyroscope that is balanced by a load in such a manner that the axis of the device is horizontal (Fig. 35). The gyroscope is rotated in the vertical plane and a load  $G$  is placed on

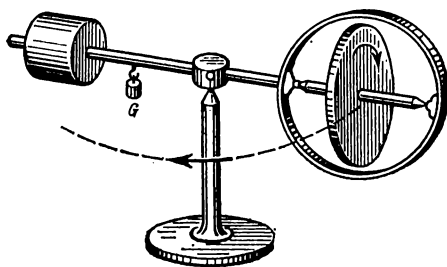


Fig. 35

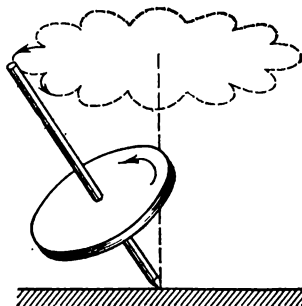


Fig. 36

the horizontal bar. It would seem that the entire right-hand portion, i.e., the gyroscope, should move upwards. Indeed, this would be the case if the gyroscope were not rotating. Actually, the rotating gyroscope begins to move with constant velocity about the vertical axis as shown by the dotted line and the arrow. This motion is at a right angle to the direction of the applied force.

This phenomenon, revolving of an axis of rotation about the direction of an applied force, is called *precession*. Everyone is familiar with the precessional motion of a top. As soon as the axis of the top begins to deviate in the least from the vertical, a gravitational torque begins acting on the top, tending to topple it. A stationary top would fall, but a rotating top begins to precess about the vertical. The axis of the top will then describe a cone whose vertex is at the point of support of the top.

In general, the rotation of a top is even more complex, for nutations are superimposed on the precessional motion. These nutations are due to small jolts (which are always present) that make the top shake (Fig. 36). As a result of the nutational effect, the axis describes a cycloidal curve, as shown in the figure, instead of a circle. It should be noted, however, that nutational effects are usually very weak.

# Vibrations

## Sec. 24. SMALL DEVIATIONS FROM EQUILIBRIUM

The motion of a body or particle about an equilibrium position is often encountered in nature. Thus, a small load on a string oscillates back and forth, a spring quivers, and an atom in a crystal lattice vibrates.

If the body or particle on which forces are acting is in a position of equilibrium, its potential energy is a minimum and the system is in a potential well (Fig. 37). When the deviation from the equilibrium position is not large, we are concerned with only a small portion of the potential well. A potential curve near the equilibrium position can always be approximated by a parabola, i.e., it can be written in the form  $U = \frac{1}{2} kx^2$ . Here,  $\frac{1}{2} k$  is the constant of proportionality. The factor  $\frac{1}{2}$  has been introduced for convenience and its purpose will presently become clear.

The reasoning used in arriving at the above relation is the following: potential energy is a function of the displacement from the equilibrium position. As is well known, making the proper assumptions, any function may be expanded in a Taylor series for small values of  $x$ . The exponent of  $x$  increases consecutively from term to term:

$$U = ax + \frac{1}{2} kx^2 + bx^3 + cx^4 + \dots$$

However, for small  $x$ , the terms of higher power may be neglected and, if the potential well is symmetrical, the first term vanishes, for the potential energies at equal distances to the left and right of equilibrium are equal.

The force acting at a point deviating from the equilibrium position is equal to minus the derivative of the potential energy. Thus, if the energy is expressed by the formula  $U = \frac{1}{2} kx^2$ , then  $F = -kx$ . The meaning of the negative sign is clear, namely, the force in question always restores the body to the equilibrium position and is always directed oppositely to the displacement. Consequently, the force  $F = -kx$  is called *the restoring force* and the coefficient  $k$  is sometimes called the restoring force constant.

What is the nature of the motion under the action of the restoring force? Newton's law, which is written in the form  $ma = -kx$  for motion near equilibrium, should give us the answer to this question.

This equation is satisfied if the point undergoes harmonic vibration about the equilibrium position, i.e., vibration in accordance with the relation

$$x = A \cos \frac{2\pi}{T} t,$$

where  $T$  is the period of vibration.

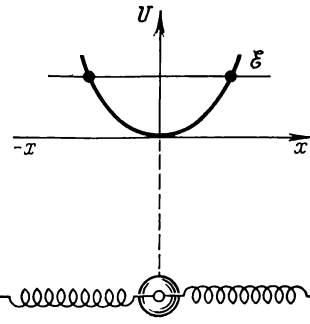


Fig. 37

Let us verify this statement. The velocity of motion of the point for the indicated dependence between displacement and time is

$$v = \frac{dx}{dt} = -\frac{2\pi A}{T} \sin \frac{2\pi}{T} t.$$

It should be noted that the maximum value for the velocity of the vibrational motion, i.e., the amplitude of the velocity, is  $v_{\max} = \frac{2\pi A}{T}$ . Let us now determine the acceleration by taking the derivative of the velocity. We obtain

$$a = -\frac{4\pi^2}{T^2} A \cos \frac{2\pi}{T} t.$$

Substituting the expressions for acceleration and displacement in Newton's law,  $ma = -kx$ , we obtain

$$-m \frac{4\pi^2}{T^2} A \cos \frac{2\pi}{T} t = -kA \cos \frac{2\pi}{T} t.$$

We see that the factors depending on time cancel out. Hence, the equation for harmonic vibrations satisfies Newton's law for small deviations from equilibrium.

It is noteworthy that Newton's law places a constraint on the period of the vibrations. As can be seen from the last formula, the period of the free vibrations about the equilibrium position is  $T = 2\pi \sqrt{\frac{m}{k}}$ . This period is determined by the vibrating system—the restoring force constant  $k$  and the mass of the particle. It is, therefore, understandable that this period is called *the natural or characteristic period* of the vibrating system.

No restrictions are placed on the amplitude  $A$  of the vibrations, with the exception, of course, that the deviations from the equilibrium position must be *small*.

## Sec. 25. PARTICULAR CASES OF VIBRATIONS

In view of the fact that we deal with two types of potential energy in mechanics, namely, elastic and gravitational, it also becomes possible to divide mechanical vibrations into these two cases.

Bodies vibrating under the action of an elastic force usually perform linear vibrations of compression and extension. However, torsional vibrations are also encountered.

If a body suspended from an elastic band, spring or wire is displaced from the equilibrium position along the band, spring or wire axis, linear vibrations arise under the action of the elastic restoring force. The coefficient  $k$  is, in this case, the stiffness of the vibrating body.

To what extent this coefficient determines the resulting period and frequency of vibration is seen from the following example. Identical loads, whose masses are equal to 1 kg, are suspended from three springs having different stiffnesses. Under the action of these loads, the springs are elongated by 1 mm, 1 cm and 1 metre, respectively. The coefficients of stiffness will then have the following values:

$$k_1 = \frac{981 \times 10^3}{0.1} = 0.981 \times 10^7 \frac{\text{dynes}}{\text{cm}};$$

$$k_2 = 0.981 \times 10^6 \frac{\text{dynes}}{\text{cm}}; \quad k_3 = 0.981 \times 10^4 \frac{\text{dynes}}{\text{cm}}.$$

The periods and frequencies of the vibrations are:

$$T_1 = 2\pi \sqrt{\frac{m}{k}} = 2\pi \sqrt{\frac{10^3}{0.981 \times 10^7}} = 6.34 \times 10^{-2} \text{ sec}, \quad \nu_1 = 15.8 \text{ Hz};$$

$$T_2 = 0.2 \text{ sec}, \quad \nu_2 = 5 \text{ Hz};$$

$$T_3 = 2 \text{ sec}, \quad \nu_3 = 0.5 \text{ Hz}.$$

For torsional vibrations, the restoration to equilibrium takes place under the action of a torsional moment that is directly proportional to the angular displacement for small deviations from equilibrium. If, for example, a massive disk having a moment of inertia  $I$  is suspended from a wire, and the wire is twisted by some angle or other, the equation for the torsional vibrations of the disk will be  $I \frac{d\omega}{dt} = -D\varphi$ . The torque  $D$ , relative to unit angular displacement, corresponds to the restoring force constant, and the moment of inertia corresponds to the mass. Thus, the period of free torsional vibrations is represented by the formula

$$T = 2\pi \sqrt{\frac{I}{D}}.$$

The greater the moment of inertia, the lower the frequency of the vibrations.

*Example.* Assume that a disk having a mass of 100 g and a radius of 5 cm is suspended from a steel wire and that the period of the torsional vibrations is 1 second. The moment of inertia of the disk is  $I_1 = \frac{mr^2}{2} = 1,250 \text{ g-cm}^2$ . Thus, the restoring force constant  $D = \frac{4\pi^2 I_1}{T^2} = 49,400 \frac{\text{dyne} \times \text{cm}}{\text{rad}}$ . If a disk of the same mass but of 1 cm radius is suspended from the same wire, the period of the torsional vibrations will no longer be 1 sec, for  $T_2 = 2\pi \sqrt{\frac{I_2}{D}} \approx 0.2 \text{ sec}$ .

A body oscillating under the action of gravitational force constitutes a pendulum. If the pendulum may be approximately represented as a point mass suspended from a weightless wire, we call it a *mathematical pendulum* (Fig. 38).

From the figure, it is easily seen that the expression for the restoring force is  $mg \sin \alpha$ , i.e., the component of the weight along the tangent to the path. If the deviation from equilibrium is small, the sine of the angle may be replaced by the value of the angle  $\alpha$  or by the quotient obtained when the displacement  $x$  is divided by the wire length  $l$ . In this approximation, displacement along the chord is assumed to coincide with displacement along the arc. Thus, the restoring force is equal to  $mg \frac{x}{l}$  and the restoring force constant is equal to  $\frac{mg}{l}$ . In the expression for the period, the mass of the bob cancels out and  $T = 2\pi \sqrt{\frac{l}{g}}$ .

The fact that the period of a pendulum does not depend on the mass is an example of a common feature of particle motion in a gravitational field. Since according to the law of gravitation the force acting on such a particle is proportional to the mass, the mass cancels out in the equation of motion. Thus, we have arrived at the well-known result that, for a given location in a gravitational field, the period of a mathematical pendulum depends only on its length.

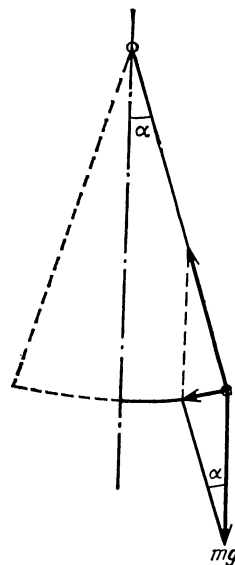


Fig. 38

The measurement of the period of a pendulum may be used to determine  $g$ . The value of this measurement may be determined extremely accurately so that very minute variations in the value of  $g$  may be ascertained. Various methods of determining the Earth's shape and various gravimetric investigations are based on this measurement. (Small changes in the value of  $g$ , which, however, greatly exceed the limits of experimental error, may occur due to seams of various density below the Earth's surface.)

When the small oscillations of a physical body cannot be approximated by a point mass, the pendulum is called a *physical pendulum*. Fig. 39 shows a rigid body whose axis of rotation (oscillation) passes through it. The period of the physical pendulum is calculated by the same formula as for torsional vibrations:

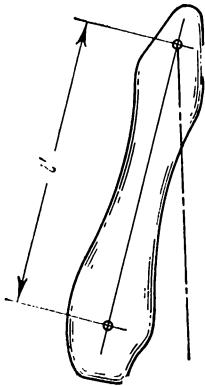


Fig. 39

$$T = 2\pi \sqrt{\frac{I}{D}},$$

since the equation

$$I \frac{d\omega}{dt} = -D\varphi$$

is valid for the motion of any body rotating about an axis. However, in the case of the gravitational field, we can easily express the torque relative to unit angular displacement by a more direct pendulum characteristic. From the same figure, it can be seen that the torque is equal to  $mgr \sin \alpha$ , i.e., the product of the weight of the body, the distance  $r$  from the centre of gravity to the point of suspension, and the sine of the angle of deviation from the equilibrium position. Since the deviation from the equilibrium position is assumed to be small—as always in this section—we obtain the expression  $mgr \alpha$  for the torque; whence,  $D = \frac{mgr \alpha}{\alpha} = mgr$ . Thus, the period of a physical pendulum is given by

$$T = 2\pi \sqrt{\frac{I}{mgr}} = 2\pi \sqrt{\frac{l'}{g}}.$$

The quantity  $l' = \frac{I}{mr}$  is called the *equivalent length* of the physical pendulum. This is the length that a mathematical pendulum would have for such a period.

## Sec. 26. TRANSFORMATION OF ENERGY. DAMPED VIBRATIONS

If there is no friction, the total energy  $\mathcal{E}$  of a body naturally remains unchanged for vibrations about its equilibrium position. Since potential energy is usually expressed relative to an arbitrary level, we shall assume that the potential energy in the equilibrium position (displacement  $x = 0$ ) is equal to zero. At any instant of motion,

$$\mathcal{E} = \frac{mv^2}{2} + \frac{kx^2}{2}.$$

In the equilibrium position, the kinetic energy is a maximum. In the end positions, the body comes to a standstill ( $v = 0$  and  $x = A$ ) and the potential energy is a maximum. It is evident from this, incidentally, that

$$\mathcal{E} = \frac{kA^2}{2},$$

i.e., the vibrational energy is proportional to the square of the amplitude.

For the three springs considered in the example on p. 66 assuming the amplitudes of the oscillations are the same, i.e.,  $A = 0.1$  cm, the total vibrational energy will have, respectively, the following values:

$$\mathcal{E}_1 = 0.49 \times 10^5 \text{ ergs}; \quad \mathcal{E}_2 = 0.49 \times 10^4 \text{ ergs}; \quad \mathcal{E}_3 = 49 \text{ ergs}.$$

This discussion has not taken into account the frictional force, which, as a rule, is experienced by all vibrating bodies. Such ideal vibrations will continue for ever without change in amplitude. Friction, however, produces damped vibrations. Formally, in this case too, it is possible to write the displacement equation in the form

$$x = A \cos \omega t,$$

but  $A$  is understood to decrease with time (Fig. 40). To determine how  $A$  depends on time, the frictional force must be known, i.e.,  $f_{fr}$  must be known for every instant of time during which vibrations occur. A simplifying assumption, more or less satisfied in practice, is that the frictional force is proportional to the velocity of motion:

$$f_{fr} = \alpha v,$$

where the coefficient  $\alpha$  is known as *the resistance constant*.

For a ball having a radius of 0.53 mm, the resistance constant  $\alpha$  at about 15°C is 13.93 g/sec in glycerine, 0.35 g/sec in sulphuric acid and 0.01 g/sec in water.

The energy equation can now be written in the form

$$d\mathcal{E} = -\alpha v dx$$

and the vibrating particle continuously loses an amount of energy equal to the work of the resisting force. Hence, the equation of motion is written as follows:

$$ma = -kx - \alpha v.$$

By substitution, it is not difficult to show that this equation is satisfied by the equation  $x = A \cos \omega t$  when the amplitude  $A$  decreases exponentially with time:

$$A = A_0 e^{-\frac{\alpha}{2m} t}.$$

Here,  $A_0$  is the amplitude at the instant of time  $t = 0$ .

It should be noted that the ratio of two successive amplitudes is a constant. Thus, the expressions for the amplitude after  $n - 1$  and  $n$  periods, respectively, are

$$A_{n-1} = A_0 e^{-\frac{\alpha}{2m} (n-1) T} \quad \text{and} \quad A_n = A_0 e^{-\frac{\alpha}{2m} n T}.$$

Let us divide the former relation by the latter. The ratio

$$\frac{A_{n-1}}{A_n} = e^{\frac{\alpha}{2m} T}$$

does not, in fact, depend on  $n$ . The rate of damping is sometimes expressed by the *logarithmic decrement*  $\delta$ :

$$\delta = \ln \frac{A_{n-1}}{A_n} = \frac{\alpha}{2m} T.$$

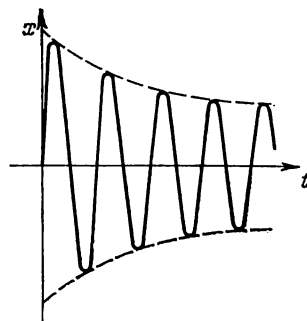


Fig. 40

Thus, the damping is greater, the greater the resistance constant, the smaller the mass, and the greater the period.

It should be noted that the period of damped vibrations differs from the period of free vibrations. The same calculation that leads to the formula for the time dependence of the amplitude also yields the following relation for the period:

$$T = T_0 \frac{1}{\sqrt{1 - \frac{\alpha^2}{4mk}}}.$$

This means that, for small resistance,  $T$  differs little from  $T_0 = 2\pi\sqrt{\frac{m}{k}}$ . When the resistance increases the period increases and, finally, for

$$\frac{\alpha^2}{4mk} = 1$$

vibrations cease. We say, in this case, that the body displaced from the equilibrium position returns aperiodically to this position.

Here are some approximate values for the logarithmic damping decrement of certain vibrating systems:

Acoustical vibrating systems . . . . .	0.1
Electrical oscillation circuits . . . . .	0.02-0.05
Tuning fork . . . . .	$10^{-3}$
Quartz crystal . . . . .	$10^{-4}$ - $10^{-5}$

Let us consider several examples of damped vibrations.

(a) Vibrating tuning fork. Logarithmic decrement  $\delta = \frac{\alpha}{2m} T = 10^{-3}$ . Assume that the period of the vibrating tuning fork is  $T = 0.01$  sec. Then,  $\frac{\alpha}{2m} = 0.1 \text{ sec}^{-1}$ . This means that during the time  $\frac{2m}{\alpha} = 10$  sec the amplitude of the vibrations decreases by the factor  $e$ :

$$A_t = A_0 e^{-\frac{\alpha}{2m} t}; \quad A_{t-10} = A_0 e^{-1}.$$

The quantity  $\frac{2m}{\alpha} = \tau$  is called *the time constant* of the given vibrating system.

(b) In acoustical vibrating systems, as can be seen from the above table, the logarithmic damping decrement is large. This means that the vibrations are rapidly damped. If  $\delta = \frac{\alpha}{2m} T = 0.1$ , the amplitude of the tenth vibration,  $A_{10}$ , will already be less than the initial amplitude  $A_0$  by the factor  $e$ . Thus,

$$\frac{A_0}{A_1} \frac{A_1}{A_2} \cdots \frac{A_8}{A_9} \frac{A_9}{A_{10}} = e^{\frac{\alpha}{2m} T \times 10}, \quad \text{i. e.,} \quad \frac{A_0}{A_{10}} = e.$$

(c) The change in the period of damped vibrations may be conveniently illustrated by means of a spring. Let a load having a mass  $m = 50$  g be suspended from a steel spring, which is thereby elongated by 2 cm. Thus, the stiffness of the spring is  $k = 24,500$  dynes/cm. If there were no damping,

$$T_0 = 2\pi \sqrt{\frac{m}{k}} = 0.28 \text{ sec.}$$



Assume that the damping is such that the time constant  $\tau_1 = \frac{2m}{\alpha} = 5$  sec, i.e., the resistance constant  $\alpha = 20$  g/sec. The period of the vibrations then becomes

$$T_1 = \frac{T_0}{\sqrt{1 - \frac{\alpha^2}{4mk}}} \approx T_0 (1 + 4.08 \times 10^{-5}).$$

Let us now immerse this pendulum in liquid. Assuming the time constant in this case to be  $\tau_2 = 1$  sec, the amplitude of the fourth vibration will already be  $1/e$  of the initial amplitude, i.e., there is considerable damping:

$$T_2 \approx T_0 (1 + 102 \times 10^{-5}) \approx 1.001 T_0.$$

Thus, even in this case the period increases by only 0.1 per cent.

## Sec. 27. FORCED VIBRATIONS

If a body is displaced from its equilibrium position and then not interfered with, the vibrations occur at the natural frequency of the body, independent of the nature of the excitation, i.e., the vibrations are determined only by the properties of the system. The frequency of the vibrations of a string remains the same regardless whether the sound was made by the string being plucked or struck.

At the same time, a number of means exist of "locking" the vibrations of a body to an external frequency. Such forced vibrations may take place if two bodies capable of vibrating are coupled. One of the bodies will force the other to vibrate. A motor that is improperly balanced will execute vibrations that are transmitted to the foundation, i.e., the foundation will execute forced vibrations. We can perform the following experiment: A pocket watch is placed in a small box and suspended by three strings. As a result, the box passes into a state of forced vibration. In Fig. 41, a device is shown in which a rotating eccentric makes a pendulum pass into a state of forced vibration. In all these cases, a periodic force varying with some frequency  $\omega$  acts on a body. Such a force is aptly called an *external force*.

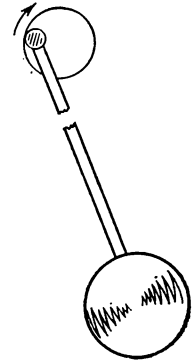


Fig. 41

Forced vibrations do not set in immediately. A certain amount of time must elapse before the body coupled to the vibrating system begins to vibrate. Eventually, a particular amplitude is reached and the frequency of the vibrations will be exactly equal to  $\omega$ .

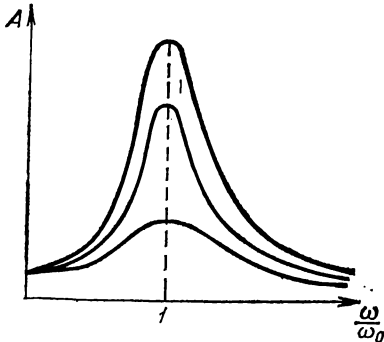


Fig. 42

The fact that a body has a natural frequency of vibration  $\omega_0$  nevertheless affects the phenomenon of forced vibrations. To be more exact, as we shall directly see, the natural frequency and the external frequency differ significantly. Fig. 42 shows the dependence of the amplitude of forced vibrations on the ratio  $\frac{\omega}{\omega_0}$  for three systems having different amounts of friction. When the external frequency and the natural frequency coincide, the amplitude of the forced vibrations is a maximum. This phenomenon is widely known as *resonance*.

The curves shown in Fig. 42 may be determined theoretically. The equation of motion of a body executing forced vibrations, under the action of a periodic exter-

nal force  $F_0 \cos \omega t$ , has the form

$$ma = -kx - \alpha v + F_0 \cos \omega t.$$

By substitution, one can easily show that the displacement of a vibrating point will satisfy the equation

$$x = A \cos (\omega t + \beta),$$

where the amplitude

$$A = \frac{F_0}{\sqrt{(\omega_0^2 - \omega^2)^2 + \alpha^2 \omega^2}}$$

and the phase shift  $\beta$  satisfies the equation

$$\tan \beta = \frac{\alpha \omega}{m(\omega_0^2 - \omega^2)}.$$

Taking into account that  $a = \frac{d^2 x}{dt^2}$  and  $v = \frac{dx}{dt}$ , let us substitute these values in the equation of motion. After simple conversion and grouping terms containing  $\cos \omega t$  and  $\sin \omega t$ , we obtain

$$[(-m\omega^2 + k) A \cos \beta - \alpha \omega A \sin \beta - F_0] \cos \omega t -$$

$$- [(-m\omega^2 + k) A \sin \beta + \alpha \omega A \cos \beta] \sin \omega t = 0.$$

Since the obtained equation must be valid for every instant of time, the coefficients of  $\cos \omega t$  and  $\sin \omega t$  must be equal to zero. Thus, we obtain two equations for determining  $A$  and  $\beta$ :

$$[(-m\omega^2 + k) \cos \beta - \alpha \omega \sin \beta] A = F_0,$$

$$[(-m\omega^2 + k) \sin \beta + \alpha \omega \cos \beta] A = 0.$$

Squaring both equations and adding, we obtain

$$A = \frac{F_0}{\sqrt{(\omega_0^2 - \omega^2)^2 + \alpha^2 \omega^2}},$$

where  $\omega_0 = \sqrt{\frac{k}{m}}$  is the frequency of the natural vibrations. From the second equation, we obtain the phase shift  $\beta$ :

$$\tan \beta = \frac{\alpha \omega}{m(\omega_0^2 - \omega^2)}.$$

From the first formula it follows that the amplitude  $A$  depends on  $\omega$  as follows: when  $\omega < \omega_0$  the amplitude increases as  $\omega$  increases; when  $\omega = \omega_0$  the amplitude reaches a maximum; and when  $\omega > \omega_0$  the amplitude decreases as  $\omega$  increases. This effect (sharpness of resonance) is more pronounced, the smaller the resistance constant  $\alpha$ . When there is little friction, the resonance disrupts the system, for at  $\alpha = 0$  the resonance amplitude goes to infinity. Engineers must take this into account in their calculations. To design a structure so that it is insensitive to the vibrations of its foundation, a resonance curve similar to the one shown in Fig. 43 must be available. The lower curve shows the vibrations of the foundation and the upper one, of the structure. At resonance, which occurs when the period of the vibrations is 0.32 sec, the amplitudes reach a value of 20-25 microns. This, in general, is no small amount.

The sharpness of resonance is an indicator of still another important phenomenon, namely, the sharper the resonance, the slower vibrations of constant amplitude set in.

Another feature of forced vibrations is the presence of phase shift. Until now, we have assumed that the origin of the coordinate system was so selected that with respect to  $t = 0$  the maximum displacement is in the positive direction. Natural-

ly, if we are considering only one vibration, there is no need to select any other origin. However, if we are comparing two vibrations and pick the origin so that  $x = A$  when  $t = 0$ , then the displacement of the other vibration at this particular instant may have an arbitrary value. This circumstance may be taken into consideration by introducing the phase shift  $\beta$  in the argument of the cosine. Thus, if  $x = A \cos(\omega t + \beta)$ , then  $x = A \cos \beta$  at the instant of time  $t = 0$ . The phase displacement is uniquely described by means of the phase shift  $\beta$ .

Let us now return to resonance phenomena. The quantity  $\beta$  in the formula for forced vibration indicates that the phase of the forced vibration, generally speaking, is shifted with respect to the phase of the impressed vibration. The magnitude of the phase shift depends on  $\frac{\omega}{\omega_0}$ , the ratio of the natural frequency to the external frequency, and also on the damping. Fig. 44 shows that a  $90^\circ$ -phase shift occurs at the resonance frequency, independent of the damping. The effect of the damping becomes clear when the situation somewhat removed from the resonance condition is considered. For weak damping (small logarithmic decrement  $\delta$ ), at frequencies somewhat below resonance, the phase shift is almost zero, while at frequencies somewhat above resonance, the phase shift is almost  $180^\circ$ . The same tendency exists for heavy

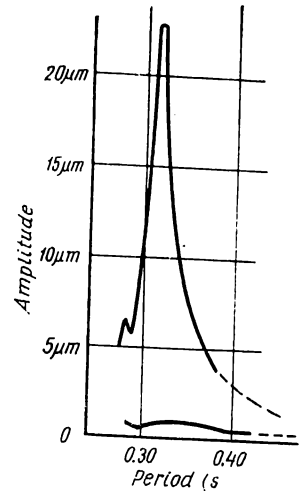


Fig. 43

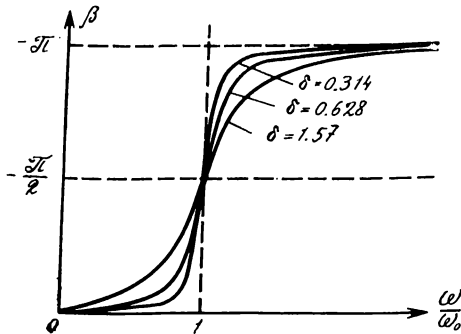


Fig. 44

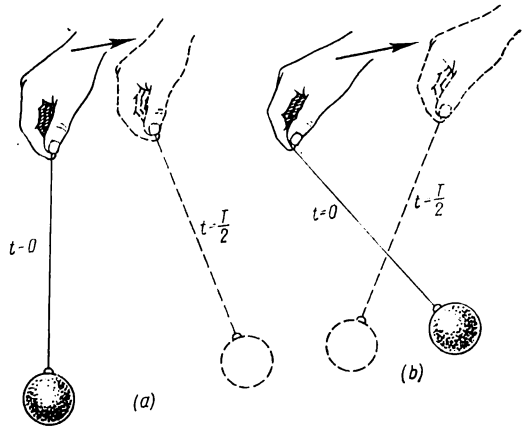


Fig. 45

damping, but it is not so pronounced. For a small amount of friction, one can say that a  $180^\circ$ -phase shift occurs when the frequency passes through the resonance condition.

A simple experiment (Fig. 45) will demonstrate the essence of these interesting relationships: Suspend a weight by a string and allow it to swing freely. When the period of the free vibrations of this pendulum is manifested, stop the pendulum and by periodic motion of the hand bring it into a state of forced vibration. At first move the hand rapidly, so that the period of the natural vibrations is greater

than the period of the forced vibrations; then move it slowly, so that the period of the natural vibrations is less than the period of the forced vibrations. It will be seen that in the first case the pendulum and the hand are  $180^\circ$  out of phase, while in the second case they are in phase.

Let us again consider the spring on page 70, which has a damping constant  $\alpha = 20$  g/sec, a mass  $m = 50$  g and a period  $T_0 = 0.28$  sec ( $\omega_0 = 22.4$  sec $^{-1}$ ). When an external sinusoidal force of frequency  $\omega = \omega_0$  acts on the spring, the amplitude of the forced vibrations is equal to  $A = 4$  cm, when the amplitude of the impressed force is  $F_0 = 1,790$  dynes  $\approx 1.8$  gf. A deviation of the frequency of the impressed force from  $\omega_0$  results in a change of the amplitude of the forced vibrations and a change in the phase shift  $\beta$  between the vibrations of the spring and the external force. The table shows the data obtained for various deviations by means of the formulas derived in this section.

Freq. of external force $\omega$ (Hz)	Amplitude of forced vibrations $A$ (cm)	Phase angle $\beta$ (degrees)
2	3.58	$0^\circ 05'$
10	3.95	$0^\circ 35'$
15	4.48	$0^\circ 15'$
22.4	4.04	$90^\circ$
30	2.48	$188^\circ 50'$
40	1.31	$189^\circ 10'$

It is seen that in the presence of damping the maximum amplitude of the forced vibrations is reached when the frequency of the impressed force is somewhat less than the natural frequency of the vibrations. The weaker the damping, the smaller this shift in frequency.

## Sec. 28. SELF-SUSTAINED VIBRATIONS

Fig. 46a shows a trough of triangular cross-section fixed on a shaft about which it can rotate. The trough has some particular period of free oscillations, which may be observed by swinging the trough away from its equilibrium position. The oscillations will continue as long as friction and air resistance do not stop them. Let us place the trough under a water faucet and allow the stream of water to flow evenly on the wall of the trough, at a point somewhat removed from the centre line. It is not difficult to envisage what will ensue. As more and more water pours into the trough, the height of the centre of gravity rises until, finally, it exceeds the height of the shaft to which the trough is fixed. The pressure of the stream of water is now sufficient to upset the trough; whereupon, water flows out and the trough returns to its original position. This cycle keeps repeating as long as the stream of water continues to flow. Thus, the trough will oscillate. However, the character of the oscillations produced in this manner is quite different from the oscillations considered above.

In the first place, it is important to note that the external force is not of an oscillatory nature; i.e., it is a constant force (the pressure of a stream of water). Secondly, such a system executes undamped oscillations, although subject to the action of friction and other resistance. And finally, the resulting oscillations are

not harmonic, i.e., they do not have a sinusoidal shape. Thus, in our example, the similarity to a sinusoid is nil. By conducting such an experiment, it can be shown that the dependence of the amount of water in the trough on the time may be represented by a saw-toothed curve similar to that shown in Fig. 46b.

The oscillations described above may be classified as *self-sustained oscillations*. Such oscillations constitute a distinct phenomenon, basically differing from free, undamped oscillations occurring without the action of a force, as well as from forced oscillations occurring under the action of a periodic force. The above example may appear to be artificial. However, self-sustained systems have broad application and are very often encountered wherever mechanical and other oscillations occur.

A simple pendulum clock (Fig. 47) executes self-sustained oscillations. As is well known, such a clock is actuated by a falling weight suspended from a chain that passes over a gear wheel. This wheel is located on the same axis as a balance wheel, which can mesh with a symmetrical anchor escapement. A pendulum is rigidly fixed to the escapement. At the instants when the balance wheel, which is driven by the gear wheel via a gear drive, touches the pallets of the escapement with its teeth, the pendulum is given an impulse. The rest of the time the pendulum and the escapement swing freely, while the balance wheel moves by itself.

The escapement and the balance wheel are so constructed that the pendulum obtains two impulses each time the balance wheel advances by one tooth. One impulse is obtained when the pendulum moves from left to right, and the other when it moves from right to left.

The self-sustained vibrations of a clock are basically similar to those of the trough of triangular cross-section. The vibrations occur under the action of a constant rather than a periodic force, are undamped in spite of the presence of friction, and are not harmonic.

In the above examples, a common property of self-sustained vibrations is manifested, namely, the property known as *feedback*. A pendulum executes undamped oscillations and causes a mechanism to give an impulse at appropriate moments. The mechanism pushes the pendulum and the pendulum provides feedback to the mechanism. If the pendulum stops, the impulses also cease. The oscillations of the pendulum are governed by the pendulum itself.

In exactly the same manner, the swings of the triangular trough are governed by the trough. The stream of water regulates the swinging of the trough, while the construction of the trough itself regulates the water inflow

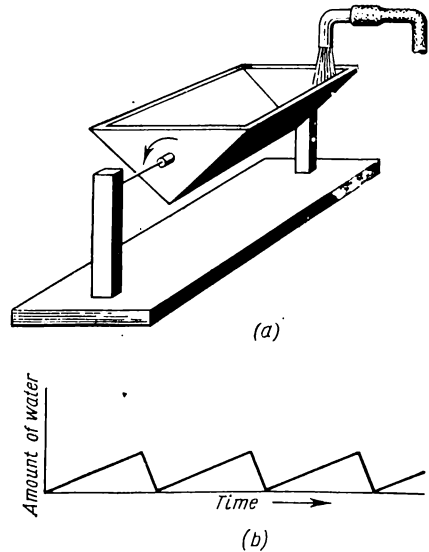


Fig. 46

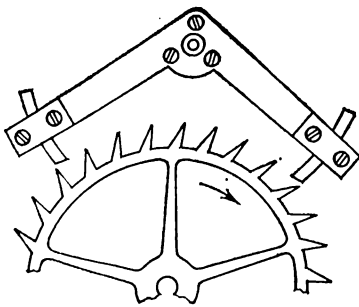


Fig. 47

A string struck with the fingers and then released is in a state of free vibration. The situation is different when a string is drawn with a bow. In this case, the string executes self-sustained vibrations that are saw-toothed in shape. The bow pulls the string along. When the displacement reaches a certain limit, the string separates away from the bow, returning to its original position. The bow again pulls the string along and the process is repeated. In the space of the second that the musician draws the bow, the phenomenon is repeated hundreds of times. These are typical self-sustained vibrations since they are due to a continuously acting force. The string itself controls the vibrations by its elasticity.

The squeaking sounds emanating from door hinges in need of oiling also belong to this class of vibrations.

We say that feedback occurs whenever an instrument or machine automatically introduces automatic corrections to its action when the operating conditions change. The principle of feedback is one of the fundamental concepts in automation.

## Sec. 29. ADDITION OF PARALLEL VIBRATIONS

In a number of cases, the problem arises of analysing the motion of a body simultaneously executing two vibrational motions. Thus, an oscillating pendulum may be located on a vibrating platform, or it may be on a rolling ship.

If we are concerned with vibrations in a single direction, the addition occurs as shown in the model in Fig. 48. Two pendulums, in this case, oscillate in parallel planes. A light rod lies freely on the pendulums and a recording pen is attached at its centre. As an approximation, we can assume that the pen will remain in a plane differing little from the planes of vibration of the pendulums and that the displacement of the pen at a given instant will be equal to the algebraic sum of the pendulum displacements. Another arrangement may also be used, e.g., a ball oscillates on a spring suspended from a board and the board, in turn, is attached to a post by a spring in such a manner that the ball simultaneously executes two different vibrations in a single plane.

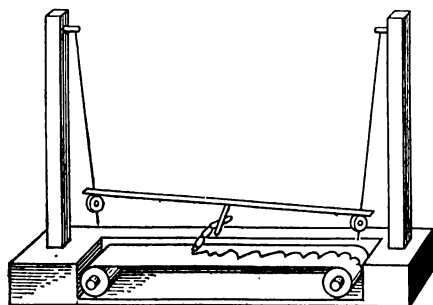


Fig. 48

If  $x_1$  is the displacement of the first vibration in the absence of the second, and  $x_2$  the displacement of the second vibration in the absence of the first, then, at each instant, for simultaneously occurring vibrational processes,

$$x = x_1 + x_2.$$

In the most general case, the component vibrations may differ in amplitude, frequency and phase.

Let us first consider the case when the vibrations have equal amplitudes and frequencies, but are displaced in phase. Then,

$$x_1 = A \cos \omega t, \quad x_2 = A \cos (\omega t + \varphi)$$

and

$$x = x_1 + x_2 = 2A \cos \frac{\varphi}{2} \cos \left( \omega t + \frac{\varphi}{2} \right),$$

where  $\omega = \frac{2\pi}{T}$ . This means that the resultant vibration is also harmonic and has the amplitude

$$2A \cos \frac{\varphi}{2}.$$

Hence, it follows that the amplitudes of vibrations add arithmetically when the vibrations coincide in phase and subtract when they are opposite in phase ( $\varphi = 180^\circ$ ). In the intermediate cases, the amplitude assumes a value between zero

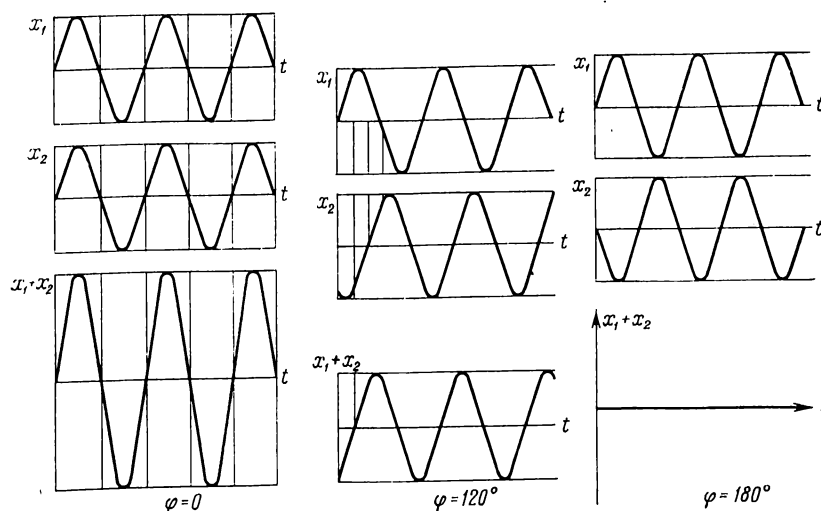


Fig. 49

and  $2A$ . In particular, when  $\varphi = 120^\circ$  the amplitude of the resultant vibration is equal to  $A$ . This is illustrated in Fig. 49.

Another important case is the addition of vibrations having different frequencies. For simplicity, let us assume that  $\varphi = 0$  and the amplitudes are equal. Then,

$$x_1 = A \cos \omega_1 t, \quad x_2 = A \cos \omega_2 t, \quad \text{and}$$

$$x = 2A \cos \frac{\omega_1 + \omega_2}{2} t \cos \frac{\omega_1 - \omega_2}{2} t.$$

In the general case, the vibrational motion obtained when such vibrations are added does not exhibit a distinct periodicity with respect to the displacement  $x$ . However, two particular cases deserve special consideration.

First, let us consider two vibrations whose frequencies are close to each other. Then,  $\omega_1 - \omega_2 \ll \omega_1 + \omega_2$  and the displacement  $x$  is the product of two cosines, one varying rapidly with time and the other very slowly. Hence,

$$2A \cos \frac{\omega_1 - \omega_2}{2} t$$

may be considered to be the slowly varying amplitude of vibrations occurring with an average frequency  $\omega_{av} = \frac{\omega_1 + \omega_2}{2}$ . The frequency of the slowly varying amplitude is known as *the beat frequency*. Fig. 50 clearly shows the two frequencies—the basic frequency of the vibrations and the beat frequency.

The second important case is the addition of two vibrations when one of the frequencies is a multiple of the other. It is quite evident that the resultant vibra-

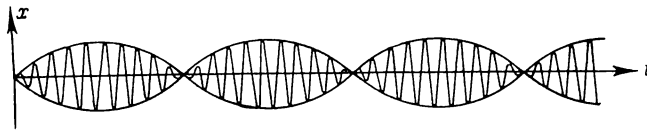


Fig. 50

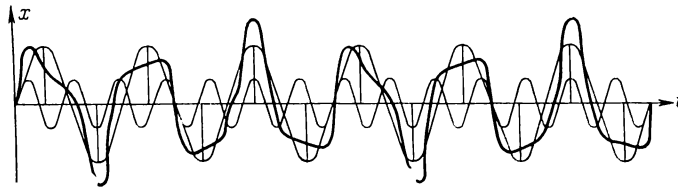


Fig. 51

tion will be periodic. If, for example, the period of one vibration is 3 sec and that of the other 7 sec, the resultant vibration will be repeated every 21 sec. This is shown in Fig. 51.

### Sec. 30. VIBRATION SPECTRUM

We have already spoken about vibrations that repeat with precision every specific interval of time, but are not harmonic. For example, we have considered saw-toothed vibrations. If we are sufficiently exacting, it turns out that harmonic vibrations, i.e., those represented by sinusoids, are encountered in nature and engineering much less often than nonharmonic vibrations.

At the end of the previous article, we noted that the sum of two sinusoids is a periodic vibration, even though not a sinusoid, if one of the frequencies is a multiple of the other. Naturally, this is true for any number of harmonic vibrations and not merely for two.

The sum of two vibrations having periods  $T$  and  $\frac{1}{2} T$ , respectively, is a vibration having a period  $T$ . Furthermore, this is the period of the vibrations obtained by adding three vibrations having periods  $T$ ,  $\frac{1}{2} T$  and  $\frac{1}{3} T$ , respectively; also, four vibrations—with the additional vibration having the period  $\frac{1}{4} T$ ; five vibrations—with the additional vibration having the period  $\frac{1}{5} T$ ; etc. Converting to frequencies, this may be expressed as follows: The sum of any number of vibrations whose frequencies are multiples of  $\omega$ , i.e., the frequencies  $\omega$ ,  $2\omega$ ,  $3\omega$ , . . . , is a vibration having the frequency  $\omega$ .

Now, the following question naturally arises: By adding an arbitrarily large number of vibrations, whose frequencies are multiples of  $\omega$  and whose amplitudes are selected as required, is it not always possible to secure any desired vibration, even saw-toothed? Fourier, the French mathematician, proved that this was indeed the case. The theorem named after him states that it is always possible to select



$a_1, a_2, a_3, \dots$  and  $\varphi_1, \varphi_2, \varphi_3, \dots$  in such a manner that any periodic vibration having the frequency  $\omega$  may be represented in the form of a sum of harmonic vibrations:

$$x = a_1 \cos(\omega t + \varphi_1) + a_2 \cos(2\omega t + \varphi_2) + a_3 \cos(3\omega t + \varphi_3) + \dots$$

The frequency  $\omega$  is called the *fundamental frequency*, and the frequencies  $2\omega, 3\omega, \dots$  are the *overtones* or *harmonics* (e.g., second harmonic, third harmonic, etc.). The closer the curve of the vibrations approaches a sinusoid, the smaller the amplitude of the harmonics. On the other hand, if the curve of the vibrations hardly resembles a sinusoid, the amplitudes of some of the harmonics will not differ greatly from the amplitude of the fundamental frequency.

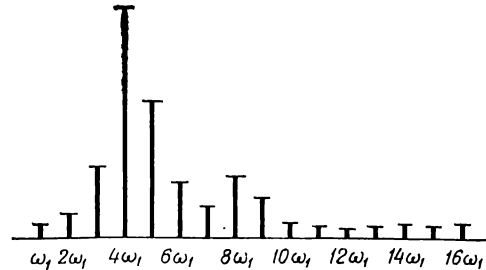


Fig. 52

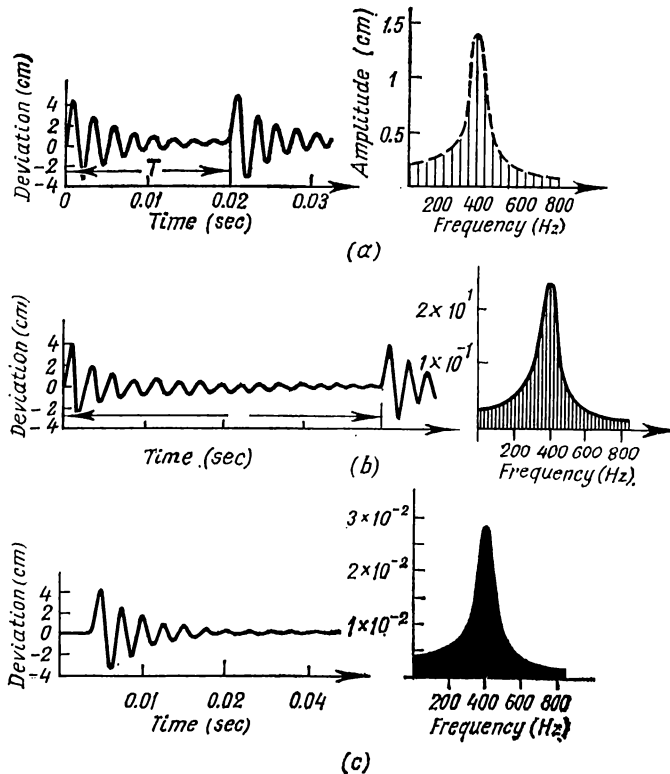


Fig. 53

Representing the vibration in the form of a sum of harmonic vibrations is called *spectrum analysis*. The spectrum consists of the data on the frequencies and amplitudes of the harmonic vibrations comprising the vibration of frequency  $\omega$ . These data may be presented in tabular form. If there are many frequencies, however, we usually resort to a graphical means of presentation (Fig. 52).

The concept of spectrum may be extended to include nonperiodic processes. We may speak of the spectrum of elastic vibrations produced by a blow on a table, the spectrum of the report of a gun, or the spectrum of an outcry.

To make this clear, let us first consider a process consisting of periodic, damped impulses. This is not the case of a single report or outcry, but of a series of reports or outcries, i.e., repeated at regular intervals of time. Characteristic of this process is the rapid damping of such a vibration, whose curve is shown in Fig. 53a. The spectrum of this vibration may be established by existing means and has the form shown on the right in the figure. As was to be expected, we see that the spectrum is composed of frequencies that are multiples of the fundamental. It should be noted that the spectrum has a maximum, which occurs for the eighth harmonic. This is not accidental, for if we examine the vibrations depicted on the left in the figure, we see that within each individual impulse the damped pulse vibrates at a "frequency" that is 8 times greater than the frequency of the fundamental tone (Fig. 53a).

Similar impulses are illustrated in Fig. 53b, but in this case the frequency is one-half of the above. Compare the spectrum of this vibration with the previous one. Since the fundamental frequency is now one-half of the original, the "frequency" of the damped, elementary process (we have assumed that it has remained the same) will now be the 16th harmonic of the fundamental tone. The distribution of the harmonic amplitudes remains as before, but their number in the same interval of frequencies is two times greater.

It is easy to see now that the spectrum of a nonperiodic process—a single impulse—is continuous. Individual frequencies are not discernible (Fig. 53c), but the nature of the spectrum is very similar to that considered above.

A mathematical proof of the above conclusions is contained in the theory of Fourier integrals.

### Sec. 31. ADDITION OF MUTUALLY PERPENDICULAR VIBRATIONS

To analyse a complex vibration consisting of the sum of two mutually perpendicular vibrations, it is best to use an electronic oscilloscope. We shall discuss this apparatus in more detail below. For the present, it is sufficient to note that an oscilloscope enables us to depict the vibrations of an electron beam in two mutually perpendicular directions. The trace of an electron beam on a fluorescent screen describes a path that is the result of two mutually perpendicular vibrational motions of the beam spot.

Let us assume that the vibration of the beam trace in the vertical direction is represented by the relation  $y = b \cos(\omega t + \delta)$ , and in the horizontal direction by the relation  $x = a \cos \omega t$ . To determine the nature of the resultant path, we must eliminate time from the above equations and obtain an equation of the form  $f(x, y) = 0$ . Writing the expressions for the displacements in the form  $\frac{x}{a} = \cos \omega t$ ,  $\frac{y}{b} = \cos(\omega t + \delta) = \cos \omega t \cos \delta - \sin \omega t \sin \delta$  and replacing, in the second equation,  $\cos \omega t$  by  $\frac{x}{a}$  and  $\sin \omega t$  by  $\sqrt{1 - \left(\frac{x}{a}\right)^2}$ , we obtain after simple conversion the equation of an ellipse rotated with respect to the coordinate axes:

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{2xy}{ab} \cos \delta = \sin^2 \delta.$$

Let us now vary the parameters of the vibrations and see what happens to the ellipse. If we vary the phase difference, the ellipse will change its form and simu-

taneously rotate (Fig. 54). When the phase difference is equal to  $90^\circ$  the axes of the ellipse coincide with the coordinate axes. If the phase difference is decreased or increased, the ellipse begins to rotate to the left or to the right, respectively, and simultaneously contracts. When the phase difference is reduced to zero, the ellipse degenerates into a straight line. The various cases can be checked by substituting, in turn, the values  $\delta = 0^\circ$ ,  $90^\circ$  and  $180^\circ$  in the above equation.

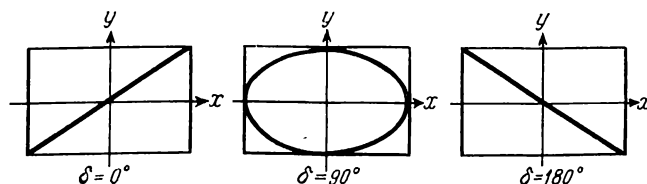


Fig. 54

If the amplitudes of the vibrations in the vertical and horizontal directions are equal, then for phase differences of  $90^\circ$  and  $270^\circ$  the path is a circle. There is a difference between these two phase differences in spite of the fact that the paths are identical. In one case, the beam moves around the circle in a clockwise direction, while in the other, the motion is counterclockwise. To see this, let us return to the original equations. We obtain the following:

$$\begin{aligned} \text{for } 90^\circ \quad x &= a \cos \omega t, \quad y = b \cos (\omega t + 90^\circ); \\ \text{for } 270^\circ \quad x &= a \cos \omega t, \quad y = b \cos (\omega t + 270^\circ). \end{aligned}$$

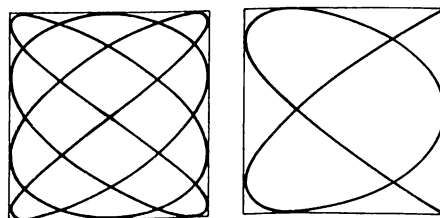


Fig. 55

The first pair of equations shows that for increasing time, at  $t = 0$ , the point having the coordinates  $x = a$ ,  $y = 0$  begins moving in the direction of negative  $y$ , i.e., clockwise. The second pair of equations shows that in this case the direction of motion is counterclockwise.

In viewing an oscillogram, it will be observed that the ellipses do not stand still, but slowly shift as though a continuous change in phase were occurring. Careful observation shows that the ellipse does not rotate, but the curve being traced by the beam spot seems to continuously shift from one ellipse to another. This phenomenon occurs when the frequencies of the vibrations differ somewhat. In fact, a difference in frequency is entirely equivalent to a continuously changing phase difference. Let us assume that the frequency of the vertical vibration  $\omega_2$  is  $\Delta\omega$  greater than the frequency of the horizontal vibration  $\omega_1$ . Then,

$$\omega_1 t + \delta = \omega_2 t + (\Delta\omega t + \delta),$$

where the variable phase difference is within the brackets.

If the frequencies differ considerably from each other, before the beam is able to describe the major portion of one ellipse the phase has already changed. As a result, the described curves look less and less like ellipses. Examples of these queer curves are shown in Fig. 55 and are known as Lissajous figures. The depicted curves are for a frequency ratio of 3 : 4 and 1 : 2.

# Travelling Waves

## Sec. 32. PROPAGATION OF A DISTURBANCE

Every body is elastic to one or another degree, i.e., every body is able to restore itself to its original form after being distorted by a force of short duration. This property is responsible for the fact that every mechanical action is transmitted by a body with finite velocity. If a perfectly rigid rod, incapable of being deformed, existed, it would only be able to move as a unit, and the action of a force would dissipate in such a body instantaneously. If a perfectly plastic body existed, deforming without in the least restoring itself to its original form, it would be incapable of transmitting any mechanical action whatsoever.

In an elastic body a disturbance is transmitted successively from one particle of a body to a contiguous one. The compression produced at the end of a rod due to the blow of a hammer is propagated along the body with a definite velocity  $c$ .

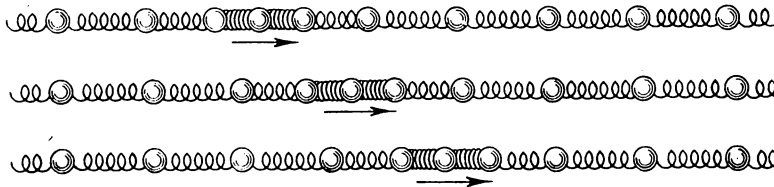


Fig. 56

If a bend of short duration is created at some point of a rigid body, this disturbance will also be transmitted with finite velocity through the body. The same is true of every deformation. Propagation through a body for various mechanical deformations is usually demonstrated by means of a spring (Fig. 56).

Elasticity of compression and extension is a property of liquid and gaseous bodies, as well as of rigid bodies. Hence, these disturbances may be transmitted in all bodies. However, the disturbances produced by shearing, torsional and bending deformations may be transmitted only by rigid bodies possessing the corresponding elasticity.

For compression and extension, the motion of the particles is in the direction in which the mechanical action is transmitted. Hence, the propagation of the disturbance is said to be longitudinal. In the case of shearing, torsion and bending, the direction of motion of the particles may form, generally speaking, any arbitrary angle with the direction in which the energy is transmitted.

We can always select the direction in which the mechanical action is being transmitted and then resolve the displacement of the particles of the body into three mutually perpendicular components, whereby one of the components is in the direction of propagation and the other two are in a perpendicular plane. Thus, in the most complex case, we can consider the propagation of a disturbance as consisting of the sum of three motions—one longitudinal and two lateral.

The velocity of propagation of the disturbance due to an elastic deformation depends on the mechanical properties of the body. As is shown in theoretical physics, this velocity is related to other physical constants of the body. Thus, for

longitudinal waves, it is given by the simple formula:

$$c = \frac{1}{\sqrt{\rho \kappa}}.$$

Here,  $\rho$  is the density of the body and  $\kappa$  is the compressibility. A large density leads to an increase in the inertia of the body's particles and, consequently, the velocity of propagation of the elastic waves decreases. The small values for compressibility indicate that even large elastic forces correspond to only small deformations. The smaller the compressibility, the greater the velocity of propagation of the disturbance.

This is the form in which this formula is generally used for liquids. Thus, water compresses by  $5 \times 10^{-5}$  of its volume for a change in pressure of 1 atm. This means that the compressibility, equal by definition to

$$\kappa = -\frac{1}{\Delta p} \frac{\Delta v}{v} \text{ (see p. 117), is } 10^{-6} \frac{\text{cm}^2}{\text{dyne}} \times 5 \times 10^{-5}.$$

The density of water is 1 g/cm<sup>3</sup>. Hence, for the velocity of propagation in water, we obtain

$$c^2 = 2 \times 10^{10} \text{ cm}^2/\text{sec}^2, \\ c = 1,400 \text{ m/sec.}$$

For gases, it is convenient to convert the formula for velocity into another form. Since the process of transmitting compression in a gas is very rapid, the compression and expansion of a gas may be considered to occur adiabatically, i.e., without heat exchange. We shall derive the equation of the adiabatic process below (p. 128), from which it is easy to obtain the following relationship between the coefficient of compressibility and the pressure of the gas:  $\kappa = \frac{1}{\gamma p}$ , where  $\gamma \approx 1.4^*$ .

Thus,  $c = \sqrt{\frac{\gamma p}{\rho}}$ . For an ideal gas, the density  $\rho = \frac{\mu}{v}$  is proportional to the fraction  $\frac{\mu}{T}$ , where  $\mu$  is the mass of a mole of gas and  $v$  is its volume. This is so because  $\frac{p v}{T} = \text{const.}$  Hence the velocity of propagation in a gas is

$$c = \sqrt{a \frac{T}{\mu}}.$$

Here,  $a$  is a constant whose value is easily calculated by means of equations considered below (p. 149).

Thus, the velocity of propagation of the disturbance due to a deformation in a gas, including the velocity of propagation of sound waves, which will be discussed in more detail later on, is proportional to the square root of the temperature

\* The equation of the adiabatic process is  $p v^\gamma = \text{const.}$  If  $p$  and  $v$  are the equilibrium values of the pressure and volume for a certain mass of gas, and  $p + \Delta p$  and  $v - \Delta v$  the corresponding values at the instant of deformation, then

$$(p + \Delta p) (v - \Delta v)^\gamma = p v^\gamma.$$

Whence,  $1 + \frac{\Delta p}{p} = \left(1 - \frac{\Delta v}{v}\right)^{-\gamma} = 1 - \gamma \frac{\Delta v}{v} + \frac{\gamma(\gamma-1)}{1 \times 2} \left(\frac{\Delta v}{v}\right)^2 + \dots$

Disregarding terms of higher order in the binomial expression, we obtain

$$\Delta p = -\gamma p \frac{\Delta v}{v}. \quad \text{Hence, } \kappa = \frac{1}{\gamma p}.$$

and does not depend on the pressure of the gas. The dependence on the molecular weight is interesting. In hydrogen, the velocity of propagation is equal to 1,263 m/sec, while in air, as is well known, it is 331 m/sec.

For longitudinal waves propagating in a rigid body, the coefficient of compressibility is usually replaced by the modulus of elasticity. Since, by definition, the modulus of elasticity

$$E = \frac{F}{S} : \frac{\Delta l}{l} = \Delta p : \frac{\Delta l}{l},$$

it is evident that in the absence of transverse motion  $\kappa = \frac{1}{E}$ , for the linear compression is equivalent to the volume compression. The formula for velocity may then be written as follows:

$$c = \sqrt{\frac{E}{\rho}}.$$

The table shows the extent of the agreement between calculated and experimental values:

	Young's Modulus $E$ , N/m <sup>2</sup>	Density $\rho$ , g/cm <sup>3</sup>	$c$ (calc.), m/sec	$c$ (exp't), m/sec
Glass . . . . .	$7.65 \times 10^{10}$	2.4	5,700	5,990
Steel . . . . .	$2.16 \times 10^{11}$	8	5,200	5,000
Wood . . . . .	$7.05 \times 10^{10}$	0.7	4,130	4,200
Water (13°C) . . . . .	$\kappa = 4.75 \times 10^{-10} \times \text{m}^2/\text{N}$	1	1,450	1,440

To check the formula for the velocity of propagation of sound, it is necessary to use samples in the shape of slim rods. This is due to the fact that a more thorough analysis of the problem indicates that the formula  $c = \sqrt{\frac{E}{\rho}}$  is only valid for such bodies. For bodies having other shapes, as well as for the propagation of sound in a continuous medium, theory leads to expression which we shall not introduce.

It should also be noted that the values given in the table are only for guiding purposes. The velocity of sound in different types of glass, wood, steel, etc., differs considerably.

### Sec. 33. GENERATION OF WAVE MOTION

Sustained vibrations may be applied to a particular point of a body or medium in a variety of ways. A force acting periodically at some point of a body produces a periodically varying deformation whose disturbance is transmitted at a specific velocity from one point of the body to another. All the particles of the body participate in the vibratory motion. Since the velocity of propagation is finite, however, the particles of the body are set into vibration in consecutive order. If a body is infinitely large, such a vibration will advance continuously, forming a travelling wave.

Infinitely large bodies do not exist. However, the actual length of a large body does not affect the nature of the phenomenon, for the vibrations do not reach the end in view of inevitable energy losses.

Let us consider a wave travelling in a particular direction in a body that for all practical purposes is infinitely large. Assume that the particle located at the origin of the coordinate system is vibrating in accordance with the equation  $y \approx$

$= A \cos \omega t$ . Let us write the equation of vibration for a particle located along the line of propagation of the disturbance at a distance  $x$  from the origin. It is not the same as for the particle at the origin because this particle began to vibrate after a delay of  $\tau = \frac{x}{c}$ , the time required for the disturbance to be propagated the distance  $x$ . The vibration of particle  $x$ , therefore, is shifted in phase with respect to the vibration of the particle at the origin. At the instant of time  $t$ , the vibration of particle  $x$  will have the same phase as the vibration of the particle at the origin at an instant of time  $\frac{x}{c}$  earlier. Hence, the equation of vibration for a particle displaced by a distance  $x$  from the origin is

$$y = A \cos \omega \left( t - \frac{x}{c} \right),$$

where  $\frac{\omega x}{c}$  is the phase shift.

The above equation is known as *the wave equation*. It is valid for the vibration of any particle located at any distance from the origin.

Let us assume that the source of the wave is far from the observer and that the wave front has long since moved ahead. We now consider a portion of the line

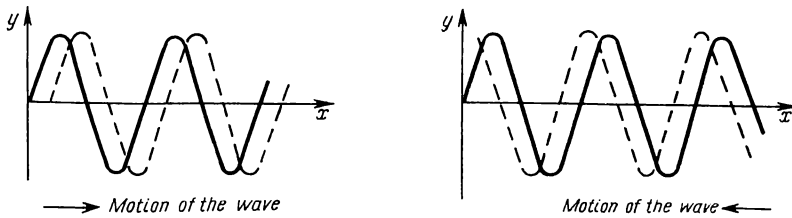


Fig. 57

along the  $x$ -axis subject to wave motion. At first glance, the introduction of a new concept may appear unjustified. To be sure, vibrations occur at all points in this region. However, can we discern how the wave is moving? Is it moving to the right or to the left? Careful observation shows that the specific character of the wave motion is easily detected. If the wave is moving from left to right, the adjacent point on the right will be delayed in phase with respect to the point on the left. In the reverse case, the point on the right will be ahead in phase. Fig. 57 shows waves travelling to the right and to the left. Each sinusoid represents a snapshot of the wave at a particular instant. At each succeeding instant, this sinusoid is displaced in its entirety in the direction in which the energy is being transmitted.

It is clear from this that the direction of the wave motion affects the form of the wave equation. If the wave moves along the coordinate axis in the positive direction, a minus sign must precede the value of the coordinate  $x$ . If the wave moves along the coordinate axis in the negative direction, the sign in the argument of the cosine must be reversed:

$$y = A \cos \omega \left( t - \frac{x}{c} \right); \quad y = A \cos \omega \left( t + \frac{x}{c} \right).$$

pos. direct.                      neg. direct.

The wave equation at an instant of time equal to a multiple of the period reduces to

$$y = A \cos \omega \frac{x}{c} = A \cos 2\pi \frac{x}{cT}.$$

The minus sign is dropped since the cosine is an even function. From this equation, it immediately follows that the period of the sinusoid is

$$\lambda = cT.$$

This spatial period, i.e., the distance covered by an undulation before it repeats itself, is known as a *wavelength*. We have thus arrived at the well-known relation connecting the velocity of a wave, its length, and the period of a vibrating particle.

A number of physical quantities vary sinusoidally in the undulatory transmission of a disturbance through a body, namely, the displacement of a particle from its equilibrium position, the velocity of the vibrating particles, the pressure and the density. Hence, the above wave equation is very general. The quantity  $y$  may designate any of the enumerated physical quantities, which vary sinusoidally when the wave moves in the  $x$ -direction. It should be noted, of course, that the waves of pressure, velocity and displacement do not necessarily have to coincide in phase. For example, it is clear that the wave representing the velocity of the vibrating particles is shifted in phase by  $90^\circ$  with respect to the wave of the displacements, for the velocity of a particle is a maximum when it passes through the equilibrium position.

#### Sec. 34. PRESSURE AND VELOCITY OF VIBRATIONS

It is interesting to examine the relationships between the wave amplitudes of various physical quantities. For this purpose, we shall only consider longitudinal waves propagating in a gas and concern ourselves with waves of displacement, particle velocity and incremental pressure. Since the theory arose in connection with auditory waves, the incremental pressure  $\Delta p$  is often called the sonic pressure and is designated by  $p$ , the symbol  $\Delta$  being dropped.

If  $A$  is the amplitude of the displacement wave, then  $\omega A$  is the amplitude of the velocity wave. These two waves are  $90^\circ$  out of phase.

We shall now derive the relationship between the amplitude of the velocity of vibrations and the amplitude of the pressure. From the general definition of  $\kappa$  as applied to gases (p. 83), we obtain for sonic pressure the formula

$$p = -\gamma P \frac{\Delta v}{v},$$

where  $P$  is the pressure of the gas. Using the relation  $c^2 = \frac{\gamma P}{\rho}$ , we obtain

$$p = -c^2 \rho \frac{\Delta v}{v}.$$

It is perfectly natural that a direct connection should exist between the incremental pressure  $p$  and the relative compression in the gas at the same location.

However, going a step further, the relative compression of the volume,  $\frac{\Delta v}{v}$ , can be related to the displacement amplitude of the vibrating particles. Let us mark two points,  $x_1$  and  $x_2$ , along the line of propagation. For a longitudinal wave, changes in density occur as a result of displacements in the direction of propagation. Let us consider a volume of gas bounded by the cross-sections through  $x_1$  and  $x_2$ . When the wave moves, the molecules within this volume are displaced. However, it is only necessary to consider the situation at the limiting cross-sections. If the molecules of the layer through  $x_1$  are displaced by  $y_1 = A \cos \omega \left( t - \frac{x_1}{c} \right)$  and



the molecules of the layer through  $x_2$  by  $y_2 = A \cos \omega \left( t - \frac{x_2}{c} \right)$ , then the linear dimension of the volume,  $x_2 - x_1$ , changes by the amount  $y_2 - y_1$ . The relative change in length, and hence in volume, is  $\frac{y_2 - y_1}{x_2 - x_1}$ . Going over to the limit, in order to obtain a quantity descriptive of a point in space, we obtain

$$\frac{\Delta v}{v} = \frac{dy}{dx} = -\frac{\omega}{c} A \sin \omega \left( t - \frac{x}{c} \right)$$

and for the pressure

$$p = c\rho A\omega \sin \omega \left( t - \frac{x}{c} \right).$$

This shows that the pressure changes in phase at the rate of the particle vibrations in the wave.  $A\omega = u_0$  is the amplitude of the velocity of vibration. Thus,  $p_0$ , the amplitude of the pressure, is related to  $u_0$ , the amplitude of the velocity, as follows:

$$p_0 = \rho c u_0.$$

In acoustics,  $u$  is usually measured in cm/sec and  $p$  in dynes/cm<sup>2</sup>. Using these units, we obtain  $p_0 = 41u_0$  for air at room temperature. The quantity  $\rho c$  is called the *acoustic* or *wave resistance*. The meaning of the designation is evident, namely, the greater the resistance the smaller the velocity of the vibrating particles for the same values of incremental pressure.

The acoustic resistance of several materials is given in the table:

	$\rho$ (g/cm <sup>3</sup> )	$c$ (cm/sec)	$\rho c$ (g/cm <sup>2</sup> sec)
Glass . . . . .	2.6	$5.5 \times 10^6$	$14 \times 10^5$
Steel . . . . .	7.9	$5 \times 10^5$	$40 \times 10^5$
Wood . . . . .	0.7	$4.2 \times 10^5$	$2.9 \times 10^5$
Water . . . . .	1	$1.44 \times 10^5$	$1.4 \times 10^5$

### Sec. 35. ENERGY FLUX

Wave motion transfers energy from one location in space to another. However, it should be kept in mind that every particle of the medium is involved in the transmission of energy and each continuously vibrates about an invariable equilibrium position.

Since all the particles of a body are involved in the vibration, the vibrational energy of a unit volume is

$$w = \frac{\rho v_{\max}^2}{2},$$

where  $\rho$  is the density, i.e., the mass of a unit volume, and  $v_{\max}$  is the amplitude of the velocity of vibration. Substituting for the latter quantity the familiar expression

$$v_{\max} = \omega A,$$

where  $A$  is the displacement amplitude and  $\omega$  is the frequency, we can write the density of the vibrational energy of a body in the form

$$w = \frac{\rho \omega^2 A^2}{2}.$$

This energy propagates with a velocity  $c$ . The following question now arises: What is the expression for the wave intensity, i.e., the amount of energy passing in unit time through a unit area perpendicular to the direction of propagation of the wave? Instead of referring to the intensity of a wave, however, it is more usual to speak about the vibrational energy flux, meaning thereby the energy passing per unit time (power) through a given area. This approach is completely analogous to the analysis of the flow of water in a pipe. In a unit time, a wave traverses a path  $c$  and transmits the energy contained in a cylinder of length  $c$

and unit cross-section. Since a unit volume contains the energy  $w$ , the energy in the above cylinder is  $wc$ . This is precisely the meaning of wave intensity:

$$I = wc.$$

We see that wave intensity has the sense of energy flow through a unit area. This was first noted by N. A. Umov in his theoretical work on energy motion in bodies.

Until now, it has been assumed that the wave motion

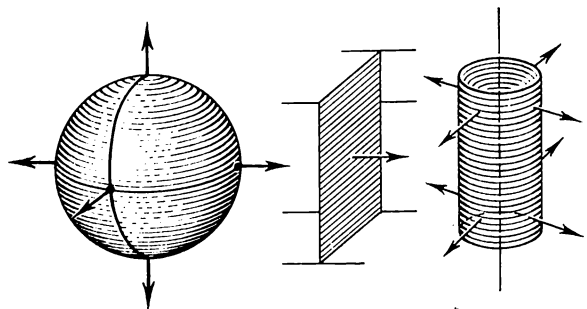


Fig. 58

propagates along a straight line. Such an assumption is useful when investigating disturbances travelling along rods, strings, air columns, etc. However, we are also interested in investigating cases involving three-dimensional wave motion.

To describe a three-dimensional wave, we must know how the wave front moves. The wave front at a particular instant can be determined if all the points in space having the same phase of vibration are given. Noting the successive positions of this constant-phase surface, i.e., the wave front, we obtain a clear picture of the nature of the wave motion.

Generally speaking, this surface may have any shape. In this case, then, what is meant by the direction of the wave propagation? It is natural to understand this direction to mean the normal to the wave front.

If the medium is perfectly homogeneous and the wave emanates from some point in the medium, the wave front is spherical. Such a wave propagates along the radii from the centre. At large distances from the centre of radiation, large portions of the wave front will appear to be in a plane—within experimental accuracy. In this manner, the concept arises of a plane wave propagating in a direction normal to the wave front. If the wave radiator is of linear shape, a cylindrical wave propagating along the radii of the cylinder is generated. Various types of waves are shown in Fig. 58.

Disregarding all energy losses that occur during the motion of a plane wave, we obtain that the quantity of energy passing through successive positions of the constant-phase surface remains unchanged. Hence, the intensity of a plane wave will not change during the process of propagation. The situation is different, however, with regard to spherical and cylindrical waves. Since the constant-phase surfaces increase in area as the square of the distance for spherical waves, and linearly as the distance for cylindrical waves, the intensities of these waves are inversely proportional to the square of the distance and to the distance, respectively. Only in this manner can the law of conservation of energy be satisfied.

The intensity of a wave is proportional to the density of the vibrational energy, which is proportional to the square of the amplitude. Hence, the amplitude

of a spherical wave is inversely proportional to the distance from the radiating centre, and the amplitude of a cylindrical wave is inversely proportional to the square root of the distance from the linear radiator. Thus,  $y = \frac{A}{r} \cos \omega \left( t - \frac{r}{c} \right)$  for a spherical wave and  $y = \frac{A}{\sqrt{r}} \cos \omega \left( t - \frac{r}{c} \right)$  for a cylindrical wave. Here, the distance  $r$ , just as previously  $x$ , is measured along the direction of wave propagation.

Let us assume that a source of vibrations having a frequency of 1 kHz is placed under water. The energy flux  $I = 1$  W/cm<sup>2</sup>. Let us calculate the displacement amplitude  $A$  of the water molecules, their acceleration  $B$  and the amplitude  $\omega A = u_0$  of the vibrational velocity.

From the formulas of previous articles, it follows that

$$A = \frac{1}{\omega} \sqrt{\frac{2I}{\rho c} \times 10^7} \text{ cm}; \quad B = \omega \sqrt{\frac{2I}{\rho c} \times 10^7} \frac{\text{cm}}{\text{sec}^2}; \quad u_0 = \sqrt{\frac{2I}{\rho c} \times 10^7} \frac{\text{cm}}{\text{sec}}.$$

For water,  $c = 1,450 \frac{\text{m}}{\text{sec}}$  and  $\rho = 1 \frac{\text{g}}{\text{cm}^3}$ . Hence,

$$A \approx 1.9 \times 10^{-3} \text{ cm}; \quad B = 740 \frac{\text{m}}{\text{sec}^2}; \quad u_0 \approx 12 \frac{\text{cm}}{\text{sec}}.$$

For the same energy flux and frequency of vibration, we obtain the following results in air, where  $c = 330 \frac{\text{m}}{\text{sec}}$  and  $\rho = 1.293 \times 10^{-3}$ :

$$A = 0.04 \text{ cm}; \quad B = 14 \times 10^5 \frac{\text{cm}}{\text{sec}^2} = 14,000 \frac{\text{m}}{\text{sec}^2}; \quad u_0 = 220 \frac{\text{cm}}{\text{sec}}.$$

### Sec. 36. DAMPING OF ELASTIC WAVES

In actuality, waves propagating in a medium (solid, liquid or gas) decrease in intensity considerably more rapidly than indicated by the inverse-square law. This is due to losses in mechanical energy, i.e., transformation of mechanical energy into heat.

The relation expressing the decrease in intensity of some particular radiation in passing through a medium may almost always be obtained by reasoning as follows (for any medium and any radiation): if a wave passes through a layer of thickness  $dx$ , the intensity loss should be proportional to the intensity of the incident wave and the thickness of the layer, i.e.,  $dI = -\mu I dx$ .

This equation may be integrated. Assuming that the intensity is equal to  $I_0$  at the point  $x = 0$  and to  $I$  at the point  $x$ , we obtain a relation that is valid for finite distances:

$$\int_{I_0}^I \frac{dI}{I} = -\mu \int_0^x dx, \quad \text{i.e.,} \quad I = I_0 e^{-\mu x}.$$

Thus, the intensity of the wave decreases exponentially.

In acoustics, it is convenient to use the concept of amplitude damping. Since the intensity is proportional to the square of the amplitude, the amplitude damping is expressed by a relation that differs from the above only in that the coefficient of damping (or absorption) is one-half of the value given there:

$$A = A_0 e^{-\frac{1}{2} \mu x}.$$

Let us examine the absorption coefficient  $\mu$  (or  $\frac{1}{2} \mu$ ) somewhat closer. It is measured in reciprocal centimetres (for the exponent must be a dimensionless quantity)

and is equal to the reciprocal of the thickness for which the intensity or amplitude of the radiation decreases by the factor  $e$ .

Naturally, the exponential damping relation is only a partial solution to the problem of the absorption of elastic waves by a medium. The determination of the dependence of the absorption coefficient on the properties of the medium and the radiation frequency is a more formidable aspect of the problem.

It has been found that for many materials the damping of an elastic wave (most of the available data being for sound waves in air) increases with the frequency of the vibration. The relation for the absorption coefficient has been determined to be

$$\mu = a\omega^2.$$

For air,  $a = 4 \times 10^{-13} \text{ sec}^2/\text{cm}$ . Thus, over a distance of 1 km, a plane wave having a frequency of 100 Hz decreases by a factor of  $\sim 1.015$ , while a very high audio frequency of 20,000 Hz decreases by a factor of  $10^{274}$ ! Ultrasonic vibrations are damped so rapidly that their transmission over a distance of more than several hundred metres is completely impractical.

However, the monotonic relationship between absorption and frequency is not always satisfied. Some materials exhibit selective absorption of sound in a relatively narrow frequency band. Thus, the absorption of ultrasonic waves by carbon dioxide is a maximum at frequencies near 277 kHz. The parabola calculated in accordance with the formula  $\mu = a\omega^2$  closely matches the experimental data in all regions except in the band indicated above. At frequencies close to 277 kHz, the absorption is about 20 times greater than that calculated assuming a parabolic relationship.

The dependence of the absorption coefficient on the properties of the medium can be expressed as follows for longitudinal waves in gases and liquids: the absorption coefficient is inversely proportional to the cube of the velocity of the elastic wave and directly proportional to the kinematic viscosity. As a result of this strong dependence on the velocity of propagation, and the fact that the kinematic viscosity of air is large, the absorption of sonic and ultrasonic waves in a liquid is about 1/1,000 of that in air. This means that for the same frequency elastic waves will propagate 1,000 times further in water than in air.

The absorption of transverse waves in solid bodies is also strongly dependent on the properties of the body. Thus, the absorption in rubber, cork and glass is, respectively, 13,000, 8,500 and 130 times greater than in aluminium.

Due to their complexity, we shall not go into the theories of elastic wave absorption in bodies.

## Sec. 37. INTERFERENCE OF WAVES

If there are several sources of waves instead of just one, then each point of the medium is simultaneously subject to several wave motions. It turns out that it is always possible to consider the vibration of a physical quantity due to the action of several waves as the sum of the vibrations that would occur if each wave acted independently.

Let us assume that spherical waves emanate from two points located at a certain distance from each other. By means of the wave equation, we can determine the value of the vibration amplitude at any instant of time for any neighbouring point. If the point in question is located at a distance  $r_1$  from the first source of waves and at a distance  $r_2$  from the second source, the vibration there is repre-

sented by the formula

$$y = A \cos 2\pi \left( vt - \frac{r_1}{\lambda} \right) + A \cos 2\pi \left( vt - \frac{r_2}{\lambda} \right).$$

The result obtained by adding two vibrations differing only in phase is, as we know, also a harmonic vibration whose amplitude is  $2A \cos \frac{\delta}{2}$ , which can be seen to depend on the phase difference between the component vibrations. The phase difference  $\delta$  is equal in this case to

$$2\pi \left( \frac{r_1 - r_2}{\lambda} \right).$$

Thus, generally speaking, all points of the wave field under consideration will be in vibration. However, the amplitudes of these vibrations will be different at different points. Two extreme cases deserve attention. First, let us consider the points at which the component vibrations annul each other. These points satisfy the condition

$$2\pi \frac{r_1 - r_2}{\lambda} = (2k + 1)\pi,$$

where  $k = 0, 1, 2, 3, \dots$ , i.e., the phase difference is equal to an odd multiple of  $\pi$ . On the other hand, if

$$2\pi \frac{r_1 - r_2}{\lambda} = 2k\pi,$$

i.e., if the phase difference is equal to an even multiple of  $\pi$ , the amplitudes of the vibrations will add arithmetically. Thus, in this case, the amplitudes reinforce each other to a maximum degree.

The difference  $r_1 - r_2$  may be called the *path difference* of the waves and this term needs no further explanation. The conditions of maximum and minimum amplitudes may be formulated somewhat differently by means of this concept. The maximum condition

$$r_1 - r_2 = k\lambda$$

states that the path difference between waves arriving at a given point must be equal to an integral number of wavelengths. The minimum condition

$$r_1 - r_2 = \frac{\lambda}{2} (2k + 1)$$

states that the path difference must be equal to an odd number of half-wavelengths. These conditions are very easily visualised as follows: The waves reinforce each other when one crest is superimposed on another, and annul each other when a crest is superimposed on a trough or node.

The superposition of waves, i.e., the addition of their amplitudes, leads to *interference*.

From analytic geometry we know that a hyperbola is a curved line satisfying the condition that the difference of the distances from any point on the curve to two foci is a constant. If we pass a plane through the point sources and note in the

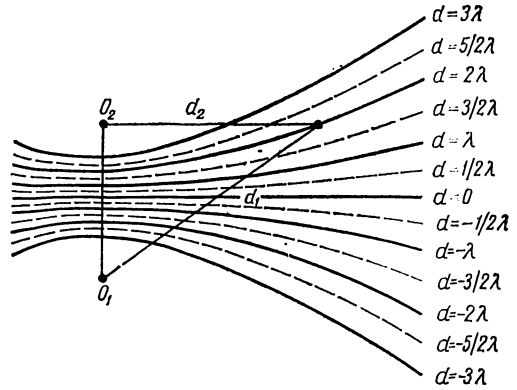


Fig. 59

diagram points of maximum reinforcement and those of wave annulment, the points will fall on hyperbolas. The corresponding curves are shown in Fig. 59. Such a picture can be easily observed on the surface of water when two sources sending out ripples from neighbouring points set up an interference pattern.

We can use the above method to analyse the interference of any number of wave sources.

### Sec. 38. PRINCIPLE OF HUYGENS-FRESNEL.

#### REFLECTION AND REFRACTION OF WAVES

The complete equality of all the vibrating points of a wave field is striking. They differ only with respect to phase. As a result, it is natural for the following idea to emerge: it should be possible to consider any point of a wave field as an independent source of spherical waves.

The validity of this idea first formulated in 1690 by Christian Huygens, may be tested by attempting to construct a wave front from data on a wave field on a boundary surface. It is necessary to take into account that there will be interference between the individual spherical waves (also called elementary waves or wavelets). Huygens' principle, supplemented by Fresnel, shows that such a construction is possible.

What is the significance of this principle? Let us assume that the wave falls on an opaque screen having several apertures. By means of the principle of Huygens-Fresnel, we can map the wave field beyond the screen without knowing anything about the sources of the field. It is sufficient to know the intensity of the field in the plane of the screen and assume that a spherical wave propagates from each point on the screen. The amplitude of the wave at any location in space is determined by adding all the wavelets coming from the apertures in the screen.

Postponing consideration of the problems related to the passage of waves through a screen (problems which are mainly of interest in connection with light waves), we shall now apply the principle of Huygens-Fresnel to the explanation of the phenomena of wave reflection and refraction.

Let us consider a portion of a plane wave incident on the boundary between two media. As is well known, a wave of any origin is reflected at an angle equal to the angle of incidence. But why should this occur? Huygens' principle gives the explanation. Every point on the boundary between the media may be considered as a wavelet source. The first wavelet emanates from the point first reached by the incident wave. Successive points on the boundary will then be excited, the last point to start vibrating being the one last reached by the incident wave. Fig. 60 shows the positions of the wavelets for the instant of time when the incident wave reaches the last point. The wave front generated by the wavelets forms an angle with the boundary equal to that of the incident wave. Thus, the propagation velocities of the incident wave and the reflected wave are the same. This means that the radius of the largest sphere must be equal to the path traversed by the incident wave from the instant the first point was excited to the instant the last point was excited.

In exactly the same manner, the wave front of a reflected spherical wave can be easily constructed. This construction is shown in Fig. 61. In Fig. 62, a photograph is shown of a sound wave reflected by a wall.

Let us now consider wavelets penetrating into the second medium and generating a refracted wave front (Fig. 63). The different media have different densities and elastic properties. Hence, the wave propagation velocity also differs for each medium. Let us now perform the same construction as for reflection, i.e., draw wave-

lets on the diagram for the instant when the incident wave reaches the last point. The wave front is rotated due to the difference in propagation velocities. If the wave has penetrated into a denser medium, the radius of the largest wavelet should be less than the path traversed by the incident wave from the instant of excitation

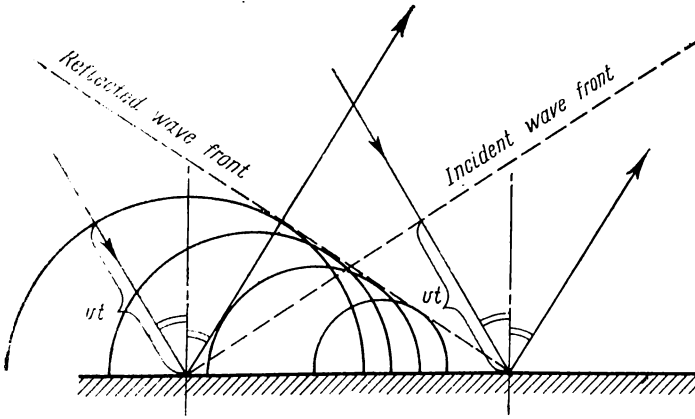


Fig. 60

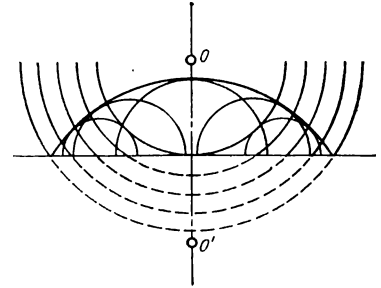


Fig. 61

of the first point on the boundary to the instant of excitation of the last point. Moreover, the ratio of these lengths should be equal to the ratio of the wave propagation velocities. On the other hand, as can be seen from Fig. 63, the ratio

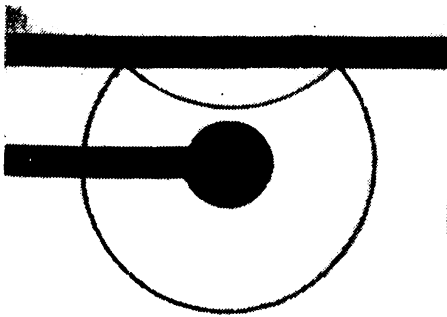


Fig. 62

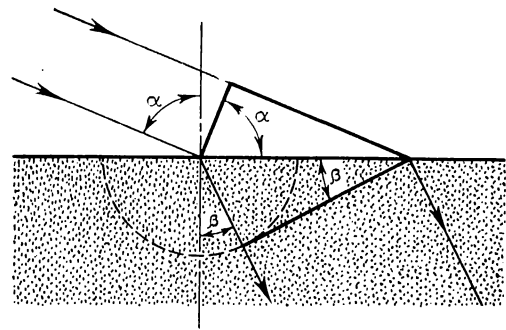


Fig. 63

of the above distances is equal to the ratio of the sine of the incident angle and the sine of the refraction angle. Thus, we arrive at the well-known relation for wave refraction:

$$\frac{\sin \alpha}{\sin \beta} = \frac{c_1}{c_2}.$$

The direction of the propagation will be deflected toward the normal to the boundary if the wave moves into a medium of greater density and, on the other hand, if it moves into a medium of lesser density, it will be deflected away from the normal. The ratio  $\frac{c_1}{c_2} = n$  is known as the *refractive index*.

## Sec. 39. REFLECTION COEFFICIENT

The explanation of the geometry of reflection and refraction may appear to be a somewhat uninteresting application of the theory. However, the wave theory enables us to do considerably more, namely, ascertain the relative proportions of the reflected and refracted waves as functions of the properties of the media whose interface is being considered. In order to simplify the calculations, we shall limit ourselves to the simple case of the normal incidence of a longitudinal wave on the interface of two media. The nature of the proof, however, is the same for all conceivable cases.

In a discussion of this type, the following is axiomatic. At the boundary between two media, neither  $u$ , the velocity of the vibrations of particles, nor  $p$ , the incremental pressure, can change abruptly. It is intuitively evident that it cannot be otherwise, but this can be shown rigorously on the basis of fundamental laws of physics.

On one side of the boundary, there are waves having the instantaneous velocities  $u_{\text{incid}}$  and  $u_{\text{reflect}}$ . On the other side of the boundary, there is also the wave having the instantaneous velocity  $u_{\text{refract}}$ . The continuity of velocity yields the condition:  $u_{\text{incid}} + u_{\text{reflect}} = u_{\text{refract}}$ ; the continuity of pressure yields:  $u_{\text{incid}}\rho_1c_1 + u_{\text{reflect}}\rho_1c_1 = u_{\text{refract}}\rho_2c_2$ . However, examining these two equations, we see that they are incompatible since  $\rho_1c_1 \neq \rho_2c_2$ . How is this to be explained? The answer is that we have forgotten that the instantaneous values of the velocities and pressures are vector quantities and that even in the simple case when the displacement vectors are in a single plane the amplitudes may differ in sign. It can be seen that the equations become compatible only if the amplitudes of the velocity and pressure for the reflected waves are given opposite signs. The equations of continuity are then written in the form

$$u_{\text{incid}} + u_{\text{reflect}} = u_{\text{refract}}; \quad (u_{\text{incid}} - u_{\text{reflect}})\rho_1c_1 = u_{\text{refract}}\rho_2c_2$$

or

$$u_{\text{incid}} - u_{\text{reflect}} = u_{\text{refract}}; \quad (u_{\text{incid}} + u_{\text{reflect}})\rho_1c_1 = u_{\text{refract}}\rho_2c_2.$$

We leave it to the reader to show that all other combinations of signs will fail to make the equations compatible.

Since the amplitudes are positive quantities, the sum must be greater than the difference. Hence, the first pair of equations is valid when  $\rho_1c_1 > \rho_2c_2$  and the second pair is valid for the reverse case. The first pair of equations arises when all the vibration velocity amplitudes are in one direction and the phase of the reflected pressure wave differs by  $180^\circ$ , i.e., the amplitude of the reflected wave is oppositely directed with respect to the incident and refracted waves. The second pair corresponds to the reverse case.

$\rho_1c_1 > \rho_2c_2$		$\rho_1c_1 < \rho_2c_2$	
Velocity wave	Pressure wave	Velocity wave	Pressure wave
incident $\rightarrow$	incident $\rightarrow$	incident $\rightarrow$	incident $\rightarrow$
reflected $\rightarrow$	reflected $\leftarrow$	reflected $\leftarrow$	reflected $\rightarrow$
refracted $\rightarrow$	refracted $\rightarrow$	refracted $\rightarrow$	refracted $\rightarrow$



This interesting phenomenon of amplitude vector reversal in reflection may be described as a one-half wavelength loss or a  $180^\circ$  phase jump. Thus, the change of sign in the wave equation  $y = A \cos \omega \left( t - \frac{x}{c} \right)$ , where  $y$  is any physical quantity, may be obtained by introducing a  $180^\circ$  phase shift in the argument of the cosine. On the other hand, a  $180^\circ$  shift in phase is equivalent to displacing the wave distribution by one-half wavelength.

Thus, at the interface of two media, the incident and reflected waves act either to reinforce each other or to annul each other to the maximum extent possible.

It should be recalled that in reflection a one-half wavelength loss occurs for the vibration velocity wave when it enters a medium of greater resistance (sometimes inaccurately expressed as a medium of greater density). The displacement wave is inseparably linked with the vibration velocity wave and also suffers a one-half wavelength loss.

In entering the second medium the wave does not execute a phase jump.

Solving the above equations simultaneously, we obtain the expression for the reflection coefficient  $r$ , i.e.,  $\frac{u_{reflect}}{u_{incident}}$ . Thus

$$r = \frac{\rho_1 c_1 - \rho_2 c_2}{\rho_1 c_1 + \rho_2 c_2},$$

where  $r$  is always  $> 0$ . Similarly, the refractive index  $g$ , i.e.,  $\frac{u_{refract}}{u_{incident}}$ , is

$$g = \frac{2\rho_1 c_1}{\rho_1 c_1 + \rho_2 c_2}.$$

The wave resistance in air is much different from that in solid bodies. As indicated above,  $\rho c = 41$  for air, while for steel ( $\rho = 7.9 \text{ g/cm}^3$  and  $c = 5,000 \text{ m/sec}$ )  $\rho c = 40 \times 10^5$ . Thus,  $r = 0.99999$ . This means that sound incident from air on steel is practically completely reflected and in effect does not penetrate the latter. It can be easily calculated that at the boundary between air and water  $r = 0.9997$ .

#### Sec. 40. THE DOPPLER EFFECT

Until now it has been assumed that the wave source and the receiver (i.e., the observer) were both stationary with respect to the medium in which the wave was propagating. Various effects, which were first noted by Doppler (1842), occur when the source or the observer, or of course both, move with respect to the medium. They consist basically in the fact that when the wave source moves the observer measures the vibration frequency  $\nu'$  and when the observer moves he measures the vibration frequency  $\nu''$ . These frequencies differ from each other and from the frequency  $\nu$  that is measured when the observer and the source are stationary.

In considering the Doppler effect, it is necessary in the first place to note that the wave leaving the source propagates entirely independently of the motion of the source and the observer. Therefore, in moving relative to the medium, the source or the observer may approach or recede from the moving wave.

Why does such motion lead to the measurement of a frequency that differs from its "real" value? This is because the observer determines the vibration frequency as the number of waves entering his apparatus per unit time. On the other hand, the formula  $\nu = \frac{c}{\lambda}$  gives the number of waves emitted per unit time. If the observer moves toward the source with the velocity  $u$ , then in 1 sec the number

of waves that he measures is not  $v$ , but a number larger than this. Moreover, the ratio of  $c + u$ , the relative velocity of the wave and the observer, to  $c$  is the factor by which the measured value is larger. Thus.

$$\frac{v'}{v} = \frac{c+u}{c}; \quad v' = v \left( 1 + \frac{u}{c} \right).$$

If the source moves toward the receiver, the observer will again register a larger number of waves than when the source and the receiver are stationary. However, in this case, the reason for the increase is different.

This is not evident at first glance. The motion of a source having a fixed frequency of vibration leads to a change in the distances between points of equal phase of the wave. If the first case is considered to be crudely analogous to the motion of an observer toward a column of athletes running at equal velocities and maintaining the constant distance  $\lambda$  between them, then the second case clearly requires a different interpretation. This case can be visualised as the slow displacement of the line of start. At equal intervals of time, the runners jump from an automobile moving along the track, which leads to a change in the distances between them. This distance becomes  $\lambda''$ , not  $\lambda'$ . If the line of start (the source) is displaced in the direction of the observer and  $v$  runners start each second, then in 1 sec they are distributed over a distance given by  $c - u$ . Thus, the interval between runners (wavelengths) is  $\lambda'' = \frac{c-u}{v}$ . The frequency at which the runners moving with velocity  $c$  cross the finish line, i.e., the vibration frequency perceived by the observer, is

$$v'' = \frac{c}{\lambda''}; \quad v'' = v \frac{1}{1 - \frac{u}{c}}.$$

Both of the above formulas are also valid when the source and the observer are moving apart. In this case, it is merely necessary to reverse the sign of  $u$ .

Thus, it has been shown that when the source and the observer approach each other the measured frequency of vibrations radiated by the source increases. When the source and the observer move apart the frequency decreases.

A well-known example of the Doppler effect is the change in sound of the whistle of a locomotive as it passes an observer. When the locomotive approaches the observer the frequency of the sound is higher than the real frequency. The pitch changes abruptly when the locomotive sweeps past him. In receding, the frequency of the sound perceived is lower than the real frequency. For a train moving at a velocity of 70 km/hr the jump amounts to  $\sim 12$  per cent of the real frequency.

# Standing Waves

## Sec. 41. SUPERPOSITION OF TWO WAVES TRAVELLING IN OPPOSITE DIRECTIONS

Let us assume that two plane waves having exactly the same characteristics are moving in opposite directions. We are interested in the resultant vibrational motion of the medium in which the waves are propagating.

As indicated above, a difference in the direction of propagation is taken into account by a difference in the coordinate signs in the wave equation. The resultant displacement should, therefore, be given by the expression

$$y = A \cos \omega \left( t - \frac{x}{c} \right) + A \cos \omega \left( t + \frac{x}{c} \right) = \\ = 2A \cos \frac{\omega x}{c} \cos \omega t = 2A \cos \frac{2\pi x}{\lambda} \cos \omega t.$$

This result is very interesting, for the sum of two travelling waves has not yielded wave motion. The formula obtained indicates the presence of vibrations of amplitude  $2A \cos \frac{2\pi x}{\lambda}$ , whose numerical value depends on the location in space. We call this peculiar vibrational state of the medium a *standing wave*, which arises whenever two identical travelling waves move in opposite directions. It should be emphasised that a standing wave is not a wave in the usual sense. A travelling wave transfers energy from one point to another, but this is in no way true of a standing wave. A travelling wave can move to the right or to the left, but a standing wave has no direction of propagation. The adopted designation merely characterises the vibrational state of the medium.

What are the characteristic features of this vibrational state? In the first place, we see that not all the particles of the medium vibrate. At the points in space satisfying the condition  $x = \frac{\lambda}{4}, \frac{3\lambda}{4}, \frac{5\lambda}{4}, \dots$ , the vibration amplitude is equal to zero. These points are known as the *nodes* of the standing wave. The distance between two adjacent nodes along the  $x$ -axis, the direction of propagation of the travelling waves, is equal to one-half wavelength. Between every two nodes is a point that vibrates with a maximum amplitude of  $2A$ . Such points are called the *antinodes* of the standing wave.

Fig. 64 shows the vibrational state corresponding to the standing wave at several successive instants of time. We see that the adopted designation is fully justified. At each instant a wave can be seen, but the wave does not move. A series of consecutive snapshots will show that the points of intersection of the wave with the abscissa, i.e., the nodes, remain fixed. The wave is stationary and the only change occurring between snapshots is a change in the magnitude of the displacements.

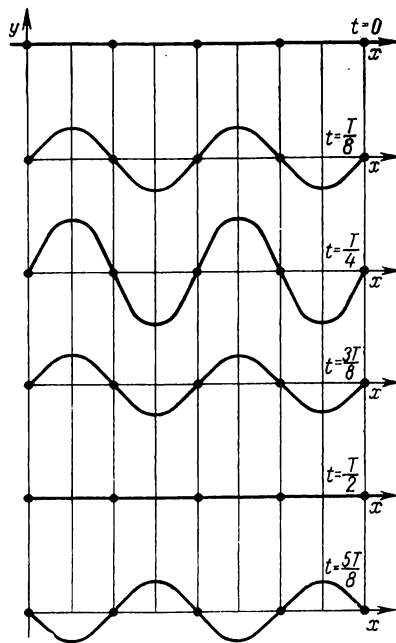


Fig. 64

At a certain instant, all points of the medium are motionless. After this instant of time, points that previously diverged upwards will now diverge downwards, and vice versa. It is evident that this picture has nothing in common with the travelling wave shown in Fig. 57, where two comparable "snapshots" are depicted. There the wave is seen to move. At each successive instant, the maxima and minima of the wave pass to new locations.

We stated that no energy transfer occurs in a standing wave. Then how can we describe, in terms of energy, the processes occurring in this peculiar vibrational motion? Clearly, the energy of a standing wave (for some region in which it exists) is a constant quantity. At the instant when all the particles are passing through the equilibrium position, all the energy of the vibrating particles is kinetic. On the other hand, at the position of maximum deviation of the particles from the equilibrium position, the energy of all the particles of the body is potential.

A standing wave is a very important vibrational process. Standing waves of different types arise in bodies of limited dimensions through which elastic waves are propagated. This is because elastic waves are reflected back from the boundary into the body. A complex vibratory state arises in the finite body which is the result of the superposition on the source wave of all the waves that were reflected from the walls. Several typical cases will now be considered.

#### Sec. 42. FREE VIBRATIONS OF A ROD

By means of a blow or other means, it is possible in any solid rod to excite a longitudinal wave that propagates along the length of the rod. From the opposite end of the rod, this wave is reflected and in this manner the entire rod is put into a vibratory state represented by a standing wave. The state will be one of free vibrations since it arises as the result of an impulse of short duration and continues without the action of external forces. We can predict the behaviour of these free vibrations if we know the length of the rod and how it is fixed. These data are known as the boundary conditions. It will be found that a node of the standing wave is located at the point where the rod is fixed and an antinode of the standing wave is located at the free end.

We shall now consider several modes of excitation of free, longitudinal vibrations in a rod of length  $L$ .

**A Rod Fixed at Both Ends.** In this case, nodes of the displacement wave are formed at the ends of the rod. Since the distance between nodes is equal to one-half wavelength, the wavelengths that are possible, in terms of the length of the rod, are given by the condition  $L = n \frac{\lambda}{2}$ , i.e.,  $\lambda_n = \frac{2L}{n}$ , where  $n$  is any integral number.

Using for the velocity of an elastic wave the expression  $c = \sqrt{\frac{E}{\rho}}$ , and recalling the relationship between frequency and wavelength, we obtain the expression for the natural frequencies of the free longitudinal vibrations of the rod:

$$\nu_n = \frac{n}{2L} \sqrt{\frac{E}{\rho}}.$$

The qualitatively new content of this result should be noted. A solid body does not have one, but a multiplicity of natural (characteristic) frequencies of vibration. Hence, a rod can execute a variety of free vibrations. It is also possible for

a rod to perform nonharmonic vibrations having any arbitrary spectrum\* consisting of the frequencies  $\nu_n$ .

The frequency  $\nu_1$  is the fundamental frequency of vibration of the rod. It corresponds to the vibratory motion for the condition  $L = \frac{\lambda}{2}$ . This means that for the fundamental vibration an antinode of the standing wave is at the centre of the rod and there are no nodes between the ends. The vibrations of the second overtone (second harmonic) correspond to the condition  $L = \lambda$ . Now, there is a node at the centre of the rod. If the third harmonic is excited, there will be two nodes between the ends of the rods, etc.

*Example.* For a steel rod ( $\rho = 7,700 \text{ kg/m}^3$  and  $E = 20.6 \times 10^{10} \text{ N/m}^2$ ) whose length is 7 metres, the fundamental frequency is  $\nu_1 = 365 \text{ Hz}$ .

**A Rod Free at Both Ends.** If a rod is suspended by a thin string and then vibrations excited in it, the resultant standing wave must satisfy the condition that antinodes are located at both ends of the rod. Just as in the previous case, the connection between the length of the rod and the wavelength is expressed by the relation:  $L = n \frac{\lambda}{2}$ . Hence, the formula for the natural frequencies will also be the same.

The difference between this case and the previous one is in the distribution of the nodes and antinodes. For the fundamental vibration, the centre of the rod is at rest (node). If the second harmonic is excited, there will be an antinode at the centre; one-quarter wavelengths away—nodes; and at the ends—antinodes.

**A Rod Fixed at One End.** In this case, there will be a node at the fixed end and an antinode at the other end. For the fundamental vibration, the rod has a form corresponding to one-quarter of a period of a sinusoid. Since the distance between a node and an antinode is equal to  $\frac{\lambda}{4}$ , the relationship between the wavelengths and the length of the rod is given by the condition

$$L = n \frac{\lambda}{4}, \quad \text{where } n = 1, 3, 5, \dots$$

The natural frequencies of the vibrations of such a rod are given by the formula

$$\nu_n = \frac{n}{4L} \sqrt{\frac{E}{\rho}} \quad (n = 1, 3, 5, \dots).$$

In the first two cases, the frequencies are related to each other as the whole numbers. Here, they are related to each other as the odd numbers.

A rod fixed at the centre will have a node at the fixed point and antinodes at the ends. The problem is essentially the same as above.

The boundary conditions used in the consideration of the vibratory state of a rod are an extreme case of the boundary conditions for reflected waves, considered on p. 94. As was explained earlier, reflection from a boundary separating one medium from another medium of greater resistance is accompanied by a loss of one-half wavelength in the displacement wave. If the rod is fixed, the wave does not penetrate the second medium at all. In this case, the second medium can be said to have an infinitely large resistance. The coefficient of reflection is equal to unity and the reflection is accompanied by a loss of one-half wavelength. It is

---

\* The word "spectrum" is used quite often in physics to denote a set of particles having different velocities, masses, etc., or a set of waves having different wavelengths (frequencies), etc.

not difficult to see that this corresponds to the presence of a node at the boundary between the two media. The reflection of the wave from the free end of the rod corresponds to reflection from a medium having zero resistance. A reflection coefficient equal to unity and the absence of a half wavelength loss leads us to conclude that an antinode must exist at such a boundary.

Free longitudinal vibrations may also be excited in columns of liquid and columns of gas.

Free lateral vibrations are easily excited in a string under tension. The distribution of the nodes and antinodes will naturally be the same as for a rod fixed at both ends. The set of frequencies is expressed by a formula analogous to that derived for the rod, the only difference being that in the expression for the velocity of the lateral wave it is necessary to replace  $E$  by the tension, i.e., the force stretching the string divided by the cross-section.

#### Sec. 43. FREE VIBRATIONS OF TWO-DIMENSIONAL AND THREE-DIMENSIONAL SYSTEMS

In rods, strings and air columns, the constant-phase surfaces consist of parallel planes. The vibratory state may be conceived as the result of superimposing plane waves extending along a single line. However, more complex cases are possible. Thus, we can have vibratory motion encompassing a two-dimensional region, e.g., plates and membranes, or encompassing a body whose three dimensions are of equal order of magnitude.

The vibration of elastic and rigid diaphragms is a two-dimensional problem. A plate fixed at its edges will have a different mode of vibration than a plate fixed at a single point or not fixed at all. Apart from the vibration of rigid plates, vibration of stretched nonrigid films, e.g., rubber and soap films, is also encountered.

In principle, the general behaviour of the free vibrations in this case does not differ from that already considered. Since this is a two-dimensional problem, the nodes and antinodes will now consist, in general, of curved lines. For example, the fundamental vibration of a circular plate fixed along its circumference has a single antinode (a point in this case) at the centre of the circle, i.e., the central point vibrates with maximum amplitude. As we move toward the edge, where a nodal circumference is located, the amplitude gradually decreases while maintaining circular symmetry. This is the simplest case, namely, vibration of the fundamental (lowest) frequency. A membrane may be excited at a higher harmonic; in such a case the surface is broken up by nodal lines. It turns out that nodal lines in a circular plate may have a circular form or consist of diameters passing through the centre.

The demonstration of nodal lines by the Chladni method (named after the scientist who proposed it) is an effective and simple experiment. A plate sprinkled with sand is put into a vibratory state by means of a blow or a fiddlestick. The sand rolls away from the antinodes and gathers along the nodal lines. Fig. 65 shows several Chladni figures.

The vibratory state of a solid three-dimensional body is, of course, the most complex. We shall avoid consideration of this phenomenon in a body of complex form and restrict our study of such free vibrations to a right-angled parallelepiped. If the standing waves in such a body were only due to the superposition of waves travelling parallel to an edge of the parallelepiped, the natural frequencies of the vibrations would be limited to the values

$$\frac{n_1 c}{2l_1}, \quad \frac{n_2 c}{2l_2}, \quad \frac{n_3 c}{2l_3},$$

and the wave numbers (the name given to the reverse of the wavelength) will be equal to

$$k_1 = \frac{n_1}{2l_1}, \quad k_2 = \frac{n_2}{2l_2}, \quad k_3 = \frac{n_3}{2l_3},$$

where  $n_1, n_2, n_3$  are arbitrary whole numbers, and  $l_1, l_2, l_3$  are the lengths of the edges of the parallelepiped.

But the waves propagating in the body may form any angle with the boundaries. The standing waves are formed only if, after a number of reflections, the beam

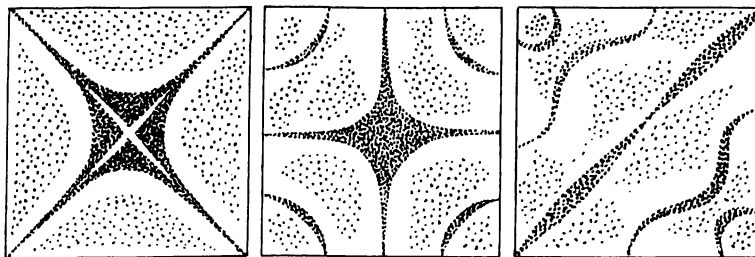


Fig. 65

returns to the exact point from which it left. The wave number of such a beam must be calculated from  $k_1, k_2, k_3$ , using the rule for vector addition. Thus,

$$k = \sqrt{k_1^2 + k_2^2 + k_3^2}, \quad \text{i. e.,} \quad v = \frac{c}{2} \sqrt{\frac{n_1^2}{l_1^2} + \frac{n_2^2}{l_2^2} + \frac{n_3^2}{l_3^2}}.$$

It is obvious that the vibration frequencies for the simple cases of wave propagation parallel to the edges of the bodies are also obtained from this formula if only one of the three whole numbers in the formula is set to be nonzero.

The vibration spectrum of a three-dimensional body is depicted in Fig. 66 in three-dimensional space, which may be called frequency space or inverse space. Here, the quantities  $\frac{c}{2l_1}, \frac{c}{2l_2}, \frac{c}{2l_3}$  are plotted, respectively,

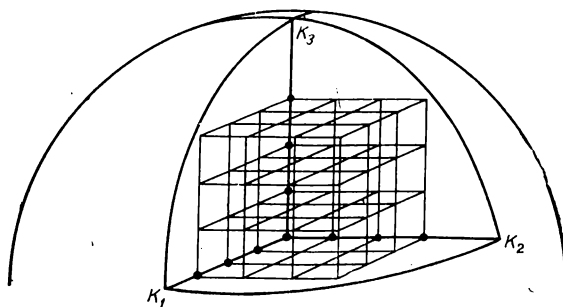


Fig. 66

Each node in the lattice (inverse lattice) thus formed represents one of the natural frequencies of vibration of the body for the numbers  $n_1, n_2, n_3$ . The radius vector drawn to a node of the lattice in the inverse space represents a possible vibration frequency. A sphere of radius  $v$  includes all points corresponding to frequencies less than  $v$ . The volume of such a sphere is equal to  $\frac{4}{3} \pi v^3$  and the volume of each cell of the inverse space is equal to  $\left(\frac{c}{2}\right)^3 / v$ , where  $v$  is the volume of the body. Therefore, the number of free vibrations of a body with frequencies less than  $v$

(the number of nodes in an octant of the sphere) is expressed by the formula

$$\frac{4}{3} \pi v \frac{v^3}{c^3}.$$

This interesting relationship shows that the number of natural frequencies increases sharply as the band of frequencies being considered increases. For high frequencies, the discrete character of the spectrum becomes blurred, for the frequencies are very close to each other.

#### Sec. 44. FORCED VIBRATIONS OF RODS AND PLATES

If the vibration of a rod, plate or other body does not take place in vacuum but in some medium\*, namely, liquid or gas, then a fraction of the intensity, depending on the ratio of the wave resistance of the contiguous media, is transferred from the vibrating body to the medium. This idea can briefly be expressed as follows:

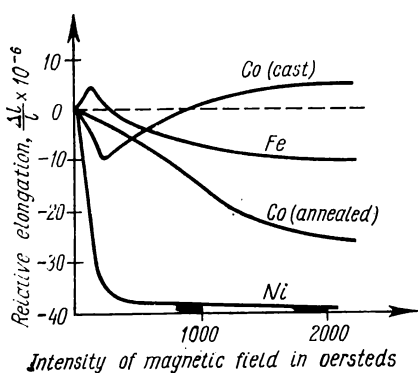


Fig. 67

a vibrating body radiates energy. Due to radiation, the free vibrations of a rod, string, etc., are rapidly attenuated. If it is required that the body be a constant source of radiation, the vibration must be excited by an external source. Just as in the case of particle vibrations, the energy may be provided either by means of self-sustained vibrations or by producing forced vibrations.

Depending on the method of providing the external energy and on its point of application, one can excite, generally speaking, one or more of the natural frequencies of a body capable of vibrating. One can, for example, produce forced

vibrations in a string under tension in the following manner. An electromagnet fed by sinusoidal current from an audio generator is fixed about a steel wire. The vibrations of the wire under the action of the periodically varying external lateral force become perceptible only at resonance. By varying the tension of the wire and the external frequency, one can show that the wire will vibrate at the fundamental frequency as well as at various overtones.

The production of forced vibrations (standing waves) in piezoelectric plates and ferromagnetic rods is of great practical importance. Such vibrating bodies are generators of ultrasonic waves.

Ferromagnetic bodies may elongate or shorten under the action of a magnetic field. The theory of this phenomenon is complex and will be treated only briefly in this book. For the present, it is sufficient to illustrate how the length of a ferromagnetic rod depends on the intensity of the field. This is done in Fig. 67, which shows that nickel and annealed cobalt shorten in fields of any intensity, cast cobalt shortens in weak fields but elongates in strong fields and, finally, iron elongates in weak fields and shortens in strong ones. In any case, a ferromagnetic rod can execute forced vibrations when placed in an alternating magnetic field. For

\* One must become reconciled to the fact that vibration of a body is used in two senses—vibration of a body as a whole and vibration of the particles of a body with respect to one another.



this purpose, the rod is usually placed in the core of a transformer fed by alternating current. In order for the standing wave in the rod to be of sufficient intensity, it is necessary to operate under conditions of resonance, i.e., the frequency of the alternating field should coincide with the rod's natural frequency of vibration. Since the rod is fixed at the centre, the natural frequency of vibrations is

$$\nu = \frac{n}{2l} c,$$

and the rod can vibrate only at the frequencies of the odd harmonics. Substituting the numerical values of the physical constants, the fundamental frequency for nickel turns out to be equal to

$$\nu = \frac{237}{l} \text{ kHz (where } l \text{ is in centimetres).}$$

Thus, a rod having a length of 40 cm will vibrate at a fundamental frequency of 6 kHz.

A piezoelectric crystal is most commonly used as a source of ultrasonic vibrations.

#### Sec. 45. PIEZOELECTRIC VIBRATIONS

Any crystal that does not have a centre of symmetry in a number of its elements of symmetry (see Sec. 274) may exhibit the piezoelectric effect. This phenomenon manifests itself in a change of the dimensions of a crystal under the action of an electric field and, conversely, in the creation of an electric field in a crystal under the action of forces applied to the crystal. When utilising the piezoelectric effect as a source of vibrations, we are dealing, of course, with the former aspect of the phenomenon, known also as *electrostriction* or *the inverse piezoelectric effect*. Piezoelectric materials include quartz crystals, Rochelle salt, barium titanate, and dihydrophosphate of ammonia. Generally speaking, there are hundreds of known materials that could, in principle, be used for this purpose. However, additional requirements, e.g., durability and stability with respect to moisture as well as the natural desire to select crystals that will yield the strongest effect, sharply limit the practical choice of material.

The change in crystal dimensions under the action of an electric field differs for different directions (with respect to the crystal's axes of symmetry). Therefore, different deformations will be obtained when, from a crystal, we cut rods or plates having different orientations with respect to the crystal's axes and place them between condenser plates. Usually, the quartz plate or other piezoelectric material is cut in such a manner that longitudinal displacements occur in the material when it is placed in an electric field. Thus, under the action of an alternating electric field, the forced vibrations produce standing longitudinal waves.

If  $l$  is the thickness of the plate in the direction of wave motion, then, as usual, the natural frequency of vibration is given by the formula  $\nu = \frac{nc}{2l}$ . For a quartz crystal having this simple orientation, the velocity of the elastic wave is equal to 5,400 m/sec. Hence, the fundamental natural frequency of vibration of a quartz plate is determined from the formula

$$\nu = \frac{2,700}{l} \text{ kHz (} l \text{ is in centimetres).}$$

It should be noted that the measured value is somewhat different, namely, 2,880/l kHz.

The vibration amplitudes depend on the magnitude of the applied field, whereby a linear dependence exists between the displacement magnitude and the electric field intensity. It is not uncommon to use large field intensities. Since quartz is an excellent insulator, electric fields of the order of 30,000 volts/cm find application for thicknesses up to a centimetre.

To attain a powerful ultrasonic signal, use is made of the resonance effect. This is essential because resonance displacements are thousands of times greater than displacements under the action of static fields, and furthermore, the vibration energy is proportional to the square of the displacement.

By gradually increasing the frequency of the generator, one can successively excite all the overtones of the crystal. The frequency range of commercial ultrasonic generators extends from hundreds to thousands of kilocycles.

## Acoustics

## Sec. 46. THE OBJECTIVE AND SUBJECTIVE NATURE OF SOUND

Man can perceive the loudness, pitch and timbre of sound by means of his hearing organs. The electronic oscilloscope enables us to investigate the objective and subjective nature of sound in detail.

Since sound is the result of a vibratory process taking place in air, it may be completely described by a curve showing amplitude change (immaterial whether displacement, vibrational velocity or pressure) with respect to time. Such a curve enables us to establish whether the process is periodic and, if so, to determine the fundamental tone of the vibration. By studying the periodicity, the overtones present and their amplitudes may be determined. In other words, the curve showing the dependence of the vibration on time always enables us to find the spectrum of the vibration, i.e., to establish which frequencies are present and the amplitudes

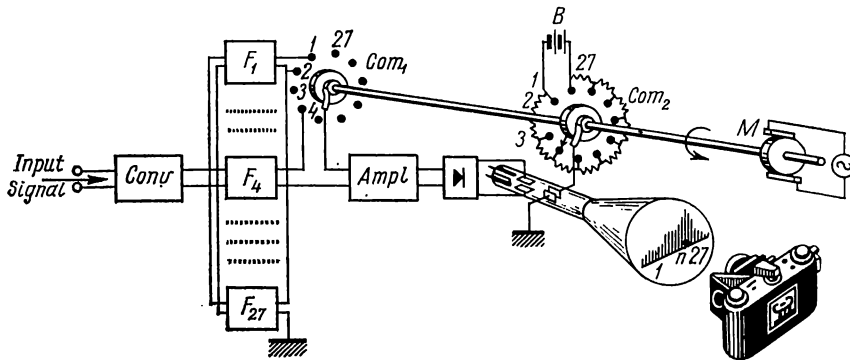


Fig. 68

of these frequencies in the spectrum. The curve is obtained by means of a microphone connected to an oscilloscope. In more elaborate arrangements, the curve of the vibration is automatically converted into its spectrum.

A simplified diagram of such an analyser is shown in Fig. 68. The input sound is converted by a microphone into electrical current, is amplified by *Conv* and applied to an apparatus consisting of a large number of filters ( $F_1, \dots, F_{27}$ ), each of which passes a specific band of frequencies, e.g.,  $\frac{1}{3}$  of an octave (36-48, 48-60, 60-72 Hz, etc.). The filters resolve the signal into its spectrum; the narrower the frequency band of each filter the greater its resolving power. Each portion of the spectrum passing through a filter is fed via a commutator  $Com_1$  to an amplifier *Ampl* and detector (rectifier). The output of the detector is applied to the oscilloscope plates that deflect the electron beam in the vertical direction. If a voltage is not applied to the second pair of oscilloscope plates, then, upon connecting each of the filters, the vertical deflection of the electron beam will be proportional to the amplitude of the corresponding frequency component of the spectrum. How-

ever, the arrangement is greatly improved by connecting a voltage to the second pair of plates that provides horizontal sweep of the electron beam. This is done by

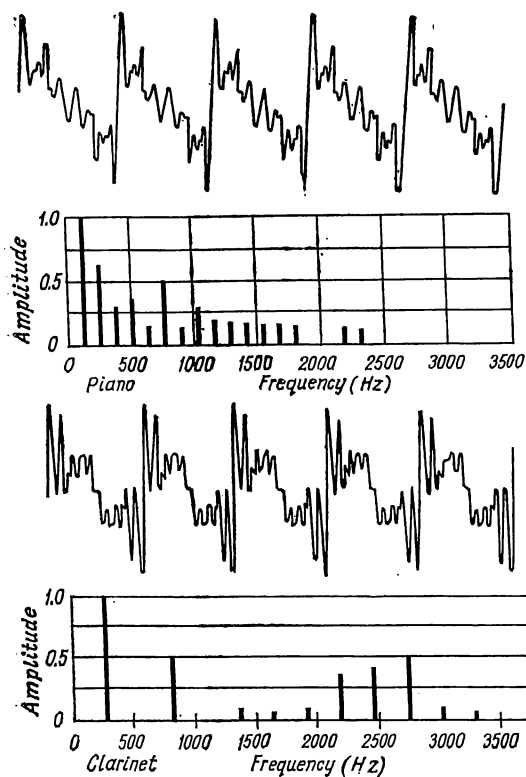


Fig. 69

properly synchronising the rotation of commutator 2, which provides the sweep voltage, with the automatic rotation of commutator 1. Thus, the amplitude of the component passed by each filter is displayed for a different, but unique, horizontal displacement of the electron beam. As a result, the complete spectrum is displayed on the oscilloscope screen.

Periodic vibrations have discrete spectra, while nonperiodic vibrations have continuous spectra. Musical sounds are illustrative of the former, while various kinds of noise illustrate the latter.

One and the same musical tone played on different instruments will have the same fundamental frequency, but will have different spectra. The quality of the sound is determined by the distribution of the overtone intensities (see Fig. 69). In the musical sense, the more complex the spectrum the richer the quality of the sound. It is interesting that the phase of the overtones (see the formula on p. 79) does not affect the subjective perception of sound. The ear can only distinguish between the intensities of the overtones.

Noise analysis is of great practical importance. If the noise frequencies of largest intensity are known, the cause of the noise is more easily determined and, consequently, more easily eliminated.

#### Sec. 47. INTENSITY AND LOUDNESS OF SOUND

In Fig. 70, the heavy lines mark the limits of the region of auditory perception for the average person. Two uniquely related quantities are plotted along the ordinate, namely, the amplitude of sound pressure and the intensity of sound. The sound pressure  $p$  and the sound intensity  $I$  are related, in the simplest case, by the formula

$$I = \frac{p^2}{2\rho c}.$$

We know that the intensity of the wave is

$$I = wc,$$

where  $w$  is the energy density, i.e.,  $w = \frac{\rho u^2}{2}$ . But  $u = \frac{p}{\rho c}$  (see p. 87). Hence, substituting, we obtain the above formula. The intensity of sound may be measured in  $\text{W/cm}^2$ .

Very intense sounds produced by a pressure of about 2,000 bars cause a sensation of pain. Very weak sounds can still be perceived by the average person when they have a pressure of  $2 \times 10^{-4}$  bar (1 bar = 1 dyne/cm<sup>2</sup>). Since for air  $\rho c = 41$ , we obtain for the limits of sound intensity the values  $0.5 \times 10^5$  ergs/sec cm<sup>2</sup> ( $= 0.5 \times 10^{-2}$  W/cm<sup>2</sup>) and  $0.5 \times 10^{-16}$  W/cm<sup>2</sup>.

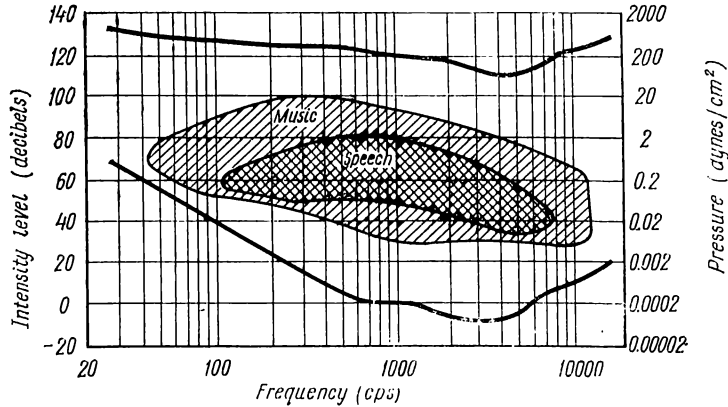


Fig. 70

This large range of intensities makes it convenient to introduce a logarithmic scale. If the intensity of one sound is  $I_1$  and the intensity of another  $I_2$ , we say that  $I_2$  is  $K$  decibels louder than  $I_1$  when

$$K = 10 \log \frac{I_2}{I_1}.$$

The quantity  $K$  is called the *loudness level*. Thus, if the sound intensities differ by a factor of a million, they differ in loudness by 60 decibels.

When expressing the sound intensity in decibels, it is necessary to indicate the zero level. The value usually selected for this level is close to the threshold of audibility ( $10^{-16}$  W/cm<sup>2</sup>). A whisper, then, has a loudness of about 15 db and the noise of an airplane about 120 db.

Returning to the diagram of auditory perception, we see that the region of speech is quite restricted in frequency (100 to 10,000 Hz) as well as in intensity (40 to 80 db). Sounds of different frequency have different audibility. The human ear perceives frequencies of several thousand cycles per second best of all. Below 20 Hz lies the infrasonic region and above 10,000-20,000 Hz the ultrasonic.

The table gives approximate values for the sound pressure  $p$ , the intensity  $I$  and the loudness  $K$ .

	$p$ (bars)	$I$ (W/cm <sup>2</sup> )	$K$ (db)
Threshold of audibility . . . . .	$2.9 \times 10^{-4}$	$10^{-16}$	0
Drip of thawing snow . . . . .	$2.9 \times 10^{-3}$	$10^{-14}$	20
Low conversation at a distance of 5 metres . . .	$2.9 \times 10^{-2}$	$10^{-12}$	40
Symphonic orchestra (fortissimo) . . . . .	2.9	$10^{-8}$	80
Airplane engine at a distance of 5 metres . . .	290	$10^{-4}$	120

## Sec. 48. ARCHITECTURE AND ACOUSTICS

In some auditoriums speech is unintelligible even though sufficiently loud, while in others the speaker must raise his voice in order to be heard. Let us investigate the physical parameters of an auditorium that determine its acoustic properties.

Experiments show that the most important factor of this nature is the so-called *reverberation time*, the time in which a sound decreases to one-millionth of its original intensity. With respect to acoustics, an auditorium is best when its reverberation time  $\tau$  is 0.5-1.5 sec. If  $\tau$  is less than 3 sec, the auditorium is considered to be good. If the reverberation time exceeds 5 sec, the acoustics are very bad, being characterised by "resounding".

A sound uttered at some part of a large hall is reflected from the walls, floor and ceiling, the furniture and drapes, and from the clothes of those present. If for each reflection the sound loses a large part of its energy, the sound will be attenuated very rapidly. The reverberation time in this case is very small and the sound will be "dull". Resounding occurs when the sound is repeatedly reflected with little attenuation. The listener will perceive the direct wave, the wave after one reflection, two reflections, etc. If the interval of time between the arrival of these sound waves does not exceed  $1/15$  of a second, the ear will not perceive two or three distinct sounds as in the case of echoes, but rather a prolonged, and hence unclear, sound.

It is evident that the time attenuation of sound is determined by its absorption in the surrounding bodies. Since the sound is repeatedly reflected, after a short time of constant sounding from some source the auditorium will more or less uniformly fill up with sonic, i.e., vibratory, energy. Within a short period of time, equilibrium is established between the energy delivered by the source and the energy absorbed by the medium. It should be noted, incidentally, that in the absence of absorption the sonic energy in a closed room would increase without limit for continuous sounding of a source.

If the sound source is interrupted, then the phenomenon reduces to the absorption of the sonic energy by the surface of the bodies located in the room. Each of the materials involved in this process has its own characteristic coefficient of absorption  $\alpha$ . If there is an open window in the room, the absorption coefficient may be assumed to be equal to 1, since the sound completely leaves the room, and this is equivalent to being absorbed. For a smooth, solid wall the coefficient  $\alpha$  is almost zero (for concrete, it is 0.015). Now, the sound absorption for the entire room may be described by the expression  $A = \alpha_1 S_1 + \alpha_2 S_2 + \alpha_3 S_3 + \dots$ , where the sum takes into account all the surfaces in the room. Theory shows that the reverberation time depends on the quantity  $A$  and the volume of the room  $V$ , i.e.,  $\tau = 0.16 \frac{V}{A}$ . In this formula, the volume is expressed in cubic metres and the quantity  $A$  in square metres.

By means of these formulas, it is not difficult to calculate the reverberation time. The absorption coefficient for concrete is given above; for glass, wood and plaster, it is not much larger (up to 3 per cent). A sharp increase in absorption occurs when soft materials are brought into the room. Suffice it to note that the clothing of one person absorbs as much sound as 20 square metres of wall surface. For soft materials, the coefficient of absorption varies between 0.5 and 0.9. A large role in the solution of acoustical problems in the construction industry is played by porous materials (e.g., spun glass and porous concrete), whose coefficients of absorption approach the values of  $\alpha$  for soft materials.

## Sec. 49. THE ATMOSPHERE AND ACOUSTICS

When a wave passes from one medium into another, it changes its direction of propagation in accordance with the law of refraction. The angle by which the direction of propagation changes is determined by the index of refraction, i.e., the ratio of the velocities of propagation.

It was indicated in Sec. 32 that the velocity of sound propagation is sensitive to changes in temperature. A temperature increase of  $1^\circ\text{C}$  increases the sound velocity by about 0.5 m/sec. The temperature of different layers of the Earth's atmosphere has, as a rule, different values. Thus, in different layers of the atmosphere, sound will have different velocities. How is the propagation of sound affected by the fact that the sound travels in a medium in which the refractive index is continuously changing?

Let us answer this question by referring to Fig. 71. Assume that the sound passes through a series of layers and that in each layer the refractive index is constant but changes abruptly from layer to layer. The path of the sound wave is represented by the broken line. If the thicknesses of the layers are small and the differences in the refractive indexes begin to decrease, the broken line will approximate a curved line. Thus, in a medium of variable index of refraction, sound waves propagate, generally speaking, along curved lines. Moreover, the path is always such that the wave travels from point to point in the shortest time. This proposition is known as Fermat's principle. A straight line, in this case, is in a certain sense not the shortest.

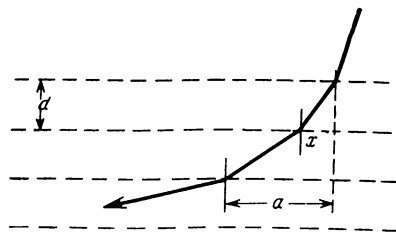


Fig. 71

We shall demonstrate the validity of this principle for the case of two adjacent segments of the broken line just considered. Let us assume, for simplicity, that the thickness  $d$  is the same for both layers and that the propagation velocities  $v_1$  and  $v_2$  are different. The time required for a wave to traverse the path indicated in the figure is equal to

$$\tau = \frac{1}{v_1} \sqrt{x^2 + d^2} + \frac{1}{v_2} \sqrt{(a-x)^2 + d^2}.$$

Here, the time is expressed in terms of the independent variable  $x$ . For different values of  $x$ , the refraction will differ and so will the time of travel from the initial point to the final point. The least time will be taken when the condition  $\frac{d\tau}{dx} = 0$  is satisfied, i.e., when

$$\frac{v_1}{v_2} = \frac{x}{\sqrt{x^2 + d^2}} : \frac{a-x}{\sqrt{(a-x)^2 + d^2}}.$$

But  $\frac{x}{\sqrt{x^2 + d^2}}$  is the sine of the incident angle and  $\frac{a-x}{\sqrt{(a-x)^2 + d^2}}$  is the sine of the refraction angle. This proves that the refraction of the wave occurs in such a manner that its time of travel is a minimum. It should be emphasised that this result is valid not only for elastic waves but for all undulatory processes.

Thus, a wave travelling in a nonhomogeneous medium changes its direction in such a manner that its path is lengthened in a medium in which the propagation velocity is larger and shortened in a medium in which the propagation velocity is smaller. In other words, a wave in a layer where the propagation velocity is large

will tend to travel parallel to the layer, while in a layer where the propagation velocity is small it will tend to travel perpendicular to the layer.

This is clearly illustrated in Fig. 72. Here, the path of a sound wave is schematically shown for the case when the temperature of the air decreases with height (usual day-time condition) and for the case when the temperature increases with height (night-time condition).

In the former case, the velocity of sound propagation is large in the layers close to the Earth. If we trace the propagation of the sound wave emanating from

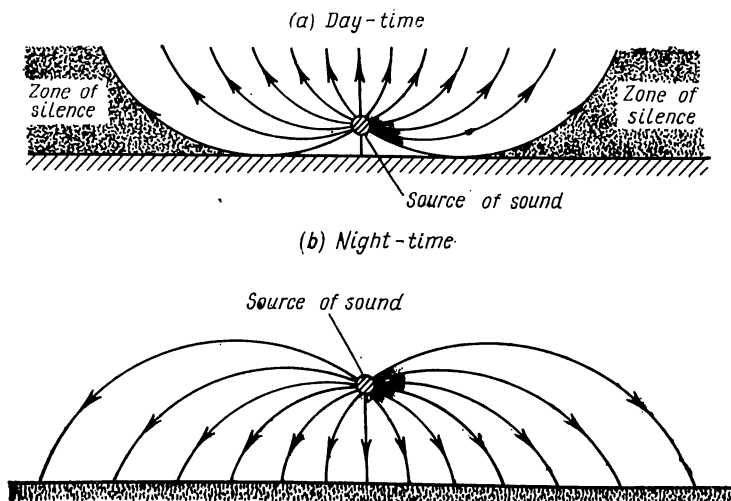


Fig. 72

a point above the Earth's surface at a small angle with the vertical, the following situation is seen to prevail. Each of the successive layers deflects the wave further and further away from the vertical. When the angle of incidence becomes equal



Fig. 73

to the angle  $i_0$ , for which  $\frac{\sin i}{n} = 1$ , refraction ceases and total reflection occurs.

Formally, the reason for total reflection is clear, namely,  $\sin i$  cannot become larger than unity. The physical basis of this interesting phenomenon will be considered below (Sec. 128) in connection with electromagnetic waves. In any case, instead of being propagated along the Earth's surface, the wave is turned in an upward direction. The diagram makes clear how "zones of silence" are formed. At night, the path of a sound wave is turned convex upwards. As a result, audibility at night is much better than during the day. When sound propagates over a reflecting surface (a still body of water), it can be heard for several kilometres even when its intensity is relatively low. The path of such a wave is represented by a series of consecutive convex arcs (Fig. 73).



## §50. ULTRASONICS

The vibratory energy in a unit volume of a sonic field is proportional to the square of the frequency. Thus, the density of the vibratory energy is  $w = \frac{\rho u^2}{2}$ , but the amplitude of the velocity is  $u_0 = A\omega$ , so that  $w$  is proportional to  $\omega^2$ . A powerful ultrasonic source is capable of producing vibrations with a pressure amplitude of dozens of atmospheres. This means that in small volumes of matter, we go through the following cycle several thousand times per second: up to dozens of atmospheres of compression, down to zero, then up to dozens of atmospheres of expansion, etc.

It is evident that a powerful mechanical action of this nature may have a number of specific effects. One such effect is *cavitation*. At the instants of vibration corresponding to maximum expansion in a liquid located in an ultrasonic field, microscopic explosions occur and dissociated gases and steam rush into this region. At the instants of vibration corresponding to compression, tremendous pressures of the order of thousands of atmospheres are produced in the regions of these explosions.

This powerful force may be used to overcome the forces acting between molecules. Emulsions such as fat in water and benzene in water become dispersed under ultrasonic action. Sooner or later cavitational explosions occur in the suspended particles. This disintegrating action has found wide application in industry.

However, ultrasonic action may be of considerable importance even when cavitation does not occur. Thus, if an ultrasonic wave is passed through aerosol (a suspension of solid particles in a gas, e.g., smoke), the particles are precipitated out. The vibrations cause the solid particles to gather at the sound pressure nodes, where the particles merge and become sufficiently heavy to fall to the ground.

Finding blowholes, internal cracks and other defects in metals by means of ultrasonic irradiation is another important field of application. The method is based on the reflection of such a wave from the boundary between the medium and air, i.e., between the metal and the inclusion. Only if the dimensions of the defect are greater than a wavelength will the method work. In order to detect a defect having a dimension of 1 mm, the wavelength should be less than 0.1 mm, i.e., a frequency of the order of  $10^9$  Hz. The frequencies used are usually much lower ( $10^7$  Hz), the method being employed to detect large flaws.

As is well known, ultrasonics also finds application in echo sounding and underwater location.

# Temperature and Heat

## Sec. 51. HEAT EQUILIBRIUM

When all the properties of a body remain unchanged, we say that the state of the body has not changed. On the other hand, when some property of the body changes, its state changes. The state of a body may be changed by doing work on it. However, the same results may be achieved without using mechanical means. Water is heated by intensely stirring it or by placing it on a gas burner. Heat exchange is said to take place when the external medium or surrounding bodies act on a body or system of bodies under consideration so as to change the state of this body or system of bodies by nonmechanical means.

If there is no heat exchange between the bodies, the bodies are in thermal equilibrium and have the same temperature. The presence of thermal equilibrium can be directly verified by bringing the bodies into contact with each other: whereupon the states of the bodies after contact should not differ from the states before contact. However, heat exchange is also possible when bodies are far apart. Thermal equilibrium may be detected, in this case, by means of a third body acting as a thermometer. If the thermometer is in equilibrium with both bodies, the temperature of these bodies is the same. This means that they would also be in a state of thermal equilibrium when in direct contact. By means of a "third body", a thermometer, it can always be ascertained whether bodies have equal or different temperatures.

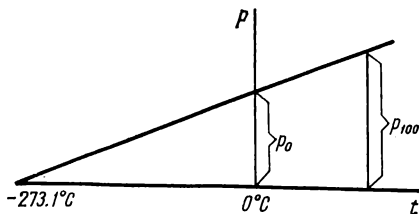


Fig. 74

By means of a thermometer, we can establish not only whether thermal equilibrium prevails or not, but also the extent of a particular deviation from equilibrium. To obtain a suitable thermometer, it is first necessary to agree on the type of thermometer (mercury, alcohol, water or gas) and the property (indication) by which we shall judge whether thermal equilibrium has been achieved between object and thermometer. As always in physics, it is important to agree on what instruments, in this case thermometers, will be considered primary. A thermometer can, then, always be calibrated by means of a standard. Gaseous hydrogen is the material used in a standard thermometer and the gas pressure  $p$  is the property by which the temperature is determined. The temperature of a body is taken as proportional to the hydrogen pressure in a gas thermometer at the constant volume occupied by the hydrogen.

The temperature scale is selected as follows. We call the temperature of the melting point of ice  $0^\circ$  and that of the boiling point of water  $100^\circ$  (at a pressure of 760 mm of mercury). Measuring the hydrogen pressures  $p_0$  and  $p_{100}$  at these two points, and drawing a straight line through the plotted points, we obtain the Celsius, or Centigrade, scale. The equation of this line, shown in Fig. 74, has the form

$$t = \frac{p - p_0}{p_{100} - p_0} \times 100.$$

The straight line intersects the  $t$ -axis at a temperature of  $-273.1^{\circ}\text{C}$ . This is absolute zero. By definition, lower temperatures are not possible. In physics, we often use a temperature calculated on the basis of absolute zero, namely,  $T = t + 273.1^{\circ}$ . This is called the absolute temperature or the temperature in degrees Kelvin ( $^{\circ}\text{K}$ ).

When we calibrate working thermometers with respect to a hydrogen standard, only a limited interval of temperatures may be used. At high temperatures, diffusion of the hydrogen through the walls of the vessel may begin to occur. At low temperatures, the hydrogen may liquefy. Nevertheless, the adopted method of determining temperature has complete general validity as will be shown below (p. 120).

## Sec. 52. INTERNAL ENERGY

The basic characteristics of bodies in the presence of mechanical and heat interaction is very well depicted by the so-called kinetic molecular model. A body consisting of molecules is considered as a system of moving and interacting particles subject to the laws of mechanics. Such a system of molecules has an energy consisting of the potential energy of the interacting particles and their kinetic energy of motion. This energy is called the *internal energy* of the body.

A specific internal energy corresponds to a specific state of the body. Changes in the mutual disposition or character of particle motion are related to changes in internal energy. Irrespective of the means employed to increase the internal energy of a body, the surrounding bodies must transfer energy to the molecules of the body under consideration. If the body is subjected to mechanical action, the energy transfer occurs in a regular manner. In the case of heat exchange, the energy is transferred through chance impulses transmitted now to one, now to another molecule.

The quantity of energy transferred to a body by mechanical means is measured by the amount of work done on the body. The quantity of energy transferred through heat exchange is measured by the quantity of heat.

Since the exact calculation of the internal energy of a body is very difficult and in most cases impossible, and since the very conception of internal energy as a purely mechanical quantity is only a rough one, a clear determination of this quantity is necessary. This can be done by considering processes occurring without heat exchange with the surroundings, i.e., so-called *adiabatic* processes. Adiabatic conditions can be provided by using a thermally insulated container for the experiment and taking the measurements during short intervals of time (so that the heat does not have time to "escape" from the volume under study). Numerous experiments leading to the establishment of the law of conservation of energy show that, irrespective of the means employed to change the state of a body in such a process, the amount of work required is exactly the same in each case. The magnitude of this work,  $A$ , is equal by definition to  $U$ , the increment of the internal energy of the body:

$$A = U_2 - U_1.$$

Naturally, the absolute value of the internal energy cannot be determined from the experiment.

If the mechanical model of a body were a completely faithful representation, the above expression would be a simple consequence of the law of conservation of mechanical energy. However, the kinetic molecular model is only a model and, therefore, the fact that there is a specific energy corresponding to each state of

a body, so that the difference in energy between two states is equal to the adiabatic work of transition, represents an extremely important law of nature leading to the law of conservation of energy.

Heat exchange and mechanical action can lead in a number of cases to the same change of state, i.e., to the same change in the internal energy of the body. This enables us to equate heat and work by measuring the quantity of heat in the same units as work and energy.

To obtain an idea of the magnitudes of various internal energies, let us cite some figures. When the temperature of 1 g of water is raised by  $1^\circ$ , the energy increases by

$$1 \text{ calorie} = 0.427 \text{ kgf-m} = 4.18 \times 10^7 \text{ ergs} = 4.18 \text{ J} = 2.61 \times 10^{19} \text{ eV}.$$

In this case, the average increase in energy of a molecule is

$$\begin{aligned} 3 \times 10^{-23} \text{ calorie} &= 1.28 \times 10^{-23} \text{ kgf-m} = 1.25 \times 10^{-15} \text{ erg} = \\ &= 12.5 \times 10^{-23} \text{ J} = 7.8 \times 10^{-4} \text{ eV}. \end{aligned}$$

The internal energy given up by matter in the combustion of 1 g of coal amounts to

$$7,000 \text{ calories} = 2,990 \text{ kgf-m} = 2.93 \times 10^{11} \text{ ergs} = 2.93 \times 10^4 \text{ J} = 18.3 \times 10^{22} \text{ eV}.$$

Calculating this on the basis of one atom of carbon, this figure reduces to

$$\begin{aligned} 1.4 \times 10^{-19} \text{ calorie} &= 5.98 \times 10^{-20} \text{ kgf-m} = 5.86 \times 10^{-12} \text{ erg} \\ &= 5.86 \times 10^{-19} \text{ J} = 3.66 \text{ eV}. \end{aligned}$$

The energy released in the nuclear fission of 1 g of uranium-235 is

$$\begin{aligned} 2.03 \times 10^{10} \text{ calories} &= 8.65 \times 10^9 \text{ kgf-m} = 8.49 \times 10^{17} \text{ ergs} \\ &= 8.49 \times 10^{10} \text{ J} = 5.29 \times 10^{29} \text{ eV} \end{aligned}$$

Calculating this on the basis of one atomic nucleus, the internal energy given up amounts to

$$\begin{aligned} 7.9 \times 10^{-12} \text{ calorie} &= 3.38 \times 10^{-12} \text{ kgf-m} = 3.3 \times 10^{-4} \text{ erg} \\ &= 3.3 \times 10^{-11} \text{ J} = 206 \times 10^6 \text{ eV} \approx 200 \text{ MeV}, \end{aligned}$$

which is more than 50 million times greater than the energy of chemical reactions ( $1 \text{ MeV} = 10^6 \text{ eV}$ ).

### Sec. 53. THE FIRST LAW OF THERMODYNAMICS

In the most general case, when energy is exchanged with the medium or with surrounding bodies, the system under consideration may receive or give up a quantity of heat  $Q$  and perform or have performed on it a given quantity of work  $A$ . Heat and work are the two forms in which the energy of a body may be transmitted to the medium or, conversely, the energy of the medium may be transmitted to the body. The law of conservation of energy excludes the possibility of any loss in the energy exchange. The difference in the energies of the system for the two states must equal the sum of the heat and work obtained by the system from the surrounding bodies.

This proposition could not be subjected to experimental verification if we did not add that the incremental energy due to the transition from one state to another is always the same irrespective of the character or method of transition from the initial to the final state. It is precisely this provision that embodies the law of conservation of energy. Now, it can clearly be subjected to all-sided experimental verification by measuring the heat and work imparted to the system for various transitions from a particular initial state to a particular final state. The incremental energy in all cases will be identical.

The law of conservation of energy expressed in the above concrete form is called *the first law of thermodynamics*. This very important law of nature was established as the result of the work of a number of scientists in the middle of the last century. The roles played by Robert Mayer, Joule and particularly Helmholtz rank especially high on the list.

In order to write the first law of thermodynamics in the form of a formula, we must first agree on the choice of sign for heat and work. Let us assume that heat is positive when it is imparted to the system and work is positive when a body performs it against the action of external forces. The first law of thermodynamics may then be written in the form

$$\Delta Q = dU + \Delta A,$$

i.e., the heat applied to a body goes to change the internal energy and perform the work of the body. Naturally, each of the quantities entering into the equation may be positive or negative depending on the particular transformation being considered.

It is not accidental that in writing the above equation the differential sign was used only for the energy. Work and heat are not total differentials. When a body goes from one state to another, the work and heat received or given up by the body depend on the "path" traversed, and only the energy increment, as in the case of the total differential of a function, does not depend on the manner of transition:

$$\int_1^2 dU = U_2 - U_1.$$

The law of conservation of energy and, in particular, the first law of thermodynamics apply in all branches of physics. Science is enabled to make particularly valuable predictions on the basis of this law. Without knowing anything about the nature of a particular process except the initial and final states of the system under consideration, a number of important conclusions may be drawn. For example, assume that molecules *A* and *B* are united by a chemical reaction to form the molecule *AB*. Assume, further, that  $U_A$ ,  $U_B$  and  $U_{AB}$ , the internal energies of the molecules, are known. If  $U_{AB}$  is greater than  $U_A + U_B$ , we are able to predict that the reaction will proceed with the absorption of heat and the quantity of heat will be equal to  $Q = U_{AB} - (U_A + U_B)$ . Or, if we know  $U_A$  and  $U_B$ , and measure the heat of the reaction by means of a calorimeter, we can determine  $U_{AB}$ . These data can then be used to predict the course of some other reaction involving the compound *AB*.

#### Sec. 54. THE INTERNAL ENERGY OF MICROSCOPIC SYSTEMS

Naturally, the law of conservation of energy and the principles of energy exchange are valid for large bodies as well as for the individual particles of a body. However, when considering particles (nuclei, atoms and molecules), or systems consisting of a small number of particles, one more important law of nature must be taken into consideration, namely, the energy of a microscopic system cannot assume any arbitrary value. Each system has a sequence of possible values of internal energy,  $E_1, E_2, \dots$ , that is characteristic of it alone. Figure 214 shows the possible energy levels for a hydrogen atom. Similar diagrams may be drawn for the energy levels of other atomic systems. When imparting heat or work to a system, the energy of the atoms, molecules or other microscopic systems may

increase only by specific, discrete amounts (quanta). In exactly the same manner, energy is given up to surrounding bodies in quanta.

Strictly speaking, the law of the quantum character of energy and the existence of a "scale" of possible energy levels for each microscopic system is a perfectly general law of nature that is valid for large bodies as well. However, as is shown in theoretical physics, the number of energy levels in a large body consisting of  $n$  atoms is, roughly speaking,  $n$  times the number of energy levels in a single atom.

As the energy increases, the intervals between levels become smaller and smaller. The reduction in the interval between energy levels is incomparable more rapid for a large body than for an individual atom and only the very lowest levels appear discrete. The higher levels merge and it appears as though a large body can change its energy continuously. If energy is taken away from a body, it "descends" to a lower level. Hence, the lower the temperature of a body, i.e., the closer it is to absolute zero, the sharper the quantum character of the energy changes.

Mechanical action serves to shift the energy levels of a body or system, but in the overwhelming majority of cases this displacement cannot be observed. For microscopic systems—atoms and molecules—the effect of pressure is very small.

Thermal interaction consists in the transitions of a system from one energy level to another.

Thermal equilibrium is mobile equilibrium. A body does not have a single energy all the time, but is rather continuously exchanging energy with its surroundings, so that on the average the energy remains unchanged. The exchange of energy occurs in discrete amounts or quanta. If at one instant the energy is equal to  $E_1$ , the next instant it has changed abruptly to  $E_2$ .

Energy is given up in the form of radiation. If  $E_1 > E_2$ , then  $E_1 - E_2 = h\nu$ , where  $\nu$  is the radiation frequency and  $h$  is Planck's constant. This constant is equal to  $6.62 \times 10^{-27}$  erg sec. Energy may be gained by absorbing radiation or as the result of a mechanical impulse from some particle.

If the temperature of a body drops instead of remaining constant, this signifies that the number of transitions from higher to lower levels exceeds the number of transitions in the reverse direction. The energy decreases in jumps and the body emits one quantum of radiation after another.

The diagrammatic representation of energy exchange was originally carried out only for atoms. Somewhat later, it became evident that this representation has general validity. We refer the reader to Part III of this book for further details on this subject.

## Sec. 55. THE EQUATION OF STATE

Three basic properties or parameters of state may be selected from the various properties of a body. These are the pressure  $p$ , the volume  $v$  and the temperature  $T$ . Knowing these parameters is not always sufficient to exhaustively describe a body. If a system consists of various substances, we must also know their concentrations. If a body is located in an electric or magnetic field, we must know the intensity of the field. However, it is always possible to select a group of parameters that will uniquely determine the state of a body. The other characteristics may then be calculated from the basic parameters.

Leaving electromagnetic fields out of consideration and restricting ourselves to simple systems—gases, liquids and isotropic solid bodies—it turns out that only two parameters determine the state of a body. It is immaterial which pair of parameters are selected from  $p$ ,  $v$  and  $T$ . Usually,  $v$  and  $T$  are selected. The pressure  $p$

is then a function of  $v$  and  $T$ . We call the equation

$$p = f(v, T)$$

*the equation of state.* It is of very great importance in physics to be able to determine this equation for a body and, in particular, for a class of bodies. Equations of state may be established only experimentally. The nature of the dependence of pressure on volume and temperature for liquids and solid bodies varies tremendously from case to case. By establishing the equation of state for a given body, we are able to determine its behaviour under a variety of conditions, but this does not in any way enable us to determine the behaviour of other bodies.

Quite often, the behaviour of a substance is not described by an equation of state, but rather by the derivatives of certain of the parameters with respect to the others.

To establish how a body expands with increasing temperature, at constant pressure, we must determine  $\left(\frac{\partial v}{\partial T}\right)_p$  (the derivative of  $v$  with respect to  $T$  at constant pressure). The quantity

$$\alpha = \frac{1}{v} \left( \frac{\partial v}{\partial T} \right)_p$$

is called *the thermometric coefficient of dilation*. As can be seen from the formula,  $\alpha$  gives the relative change in the volume of a body for a  $1^\circ$  change in temperature.

*The thermometric coefficient of change of pressure*

$$\beta = \frac{1}{p} \left( \frac{\partial p}{\partial T} \right)_v$$

gives the relative change in the pressure for a  $1^\circ$  change in temperature (at constant volume). The dimension of the coefficients  $\alpha$  and  $\beta$  is reciprocal to the degree, i.e.,  $K^{-1}$ .

The third useful quantity is *the compressibility*

$$\kappa = -\frac{1}{v} \left( \frac{\partial v}{\partial p} \right)_T,$$

which gives the relative decrease in volume for a unit increase in pressure (at constant temperature).

These three coefficients are connected by a very interesting relationship, which is easily derived as follows: Since

$$p = f(v, T),$$

then

$$dp = \left( \frac{\partial p}{\partial T} \right)_v dT + \left( \frac{\partial p}{\partial v} \right)_T dv.$$

If the pressure is constant, then  $dp = 0$  and

$$\left( \frac{\partial p}{\partial T} \right)_v \left( \frac{\partial T}{\partial v} \right)_p \left( \frac{\partial v}{\partial p} \right)_T = -1.$$

Hence

$$\frac{\beta \kappa}{\alpha} = \frac{1}{p}.$$

This result shows that if we know, for example, the compressibility and the thermometric coefficient of change of pressure, we can calculate the value of the thermometric coefficient of dilation. The derived relationship is valid for all bodies.

Generally speaking, the coefficients  $\alpha$ ,  $\beta$  and  $\kappa$  are not constant quantities for a given substance. For various pressures and temperatures, these coefficients may assume various values. Hence, when the value of a coefficient is given, it is necessary to indicate the corresponding values of pressure and temperature. In some cases, the average value of the coefficients is given over a particular interval of temperatures or pressures.

Here are some examples:

(a) The table gives the thermometric coefficient of dilation  $\alpha$  and the compressibility  $\kappa$  for some liquids.

	$\alpha$ , K <sup>-1</sup>	$\kappa$ , m <sup>2</sup> /N
Water, 10°—30° C, normal pressure . . . . .	$2.07 \times 10^{-4}$	$48.5 \times 10^{-11}$
Mercury, 10°—30° C . . . . .	$1.81 \times 10^{-4}$	$3.05 \times 10^{-11}$
Ether, 0° C . . . . .	$16.56 \times 10^{-4}$	$149 \times 10^{-11}$

For solid bodies, the thermometric coefficient of dilation and the compressibility may vary considerably. Thus, at normal temperature and pressure, for fused quartz,  $\alpha = 1.29 \times 10^{-6}$  K<sup>-1</sup>, and  $\kappa = 2.76 \times 10^{-11}$  m<sup>2</sup>/N, while for ebonite  $\alpha = 77 \times 10^{-6}$  K<sup>-1</sup>, and  $\kappa = 18.4 \times 10^{-11}$  m<sup>2</sup>/N.

(b) The values given in the table for  $\beta$ , the thermometric coefficient of change of pressure, are calculated for water, mercury and ether at atmospheric pressure ( $\frac{\beta\kappa}{\alpha} = 1$ ).

	Water	Mercury	Ether
$\beta$ , K <sup>-1</sup> . . . . .	4.4	61.4	11.3

This means that when the temperature of a *constant* volume of mercury is increased by 10<sup>-3</sup> degree, its pressure increases by 6 per cent (!).

## Sec. 56. THE EQUATION OF THE GAS STATE

The simplest equation of state is that of a rarefied gas. It was obtained by Mendeleev in the form of a single formula that combines Clapeyron's equation and Avogadro's law. Clapeyron's equation states that  $\frac{pv}{T}$  is a constant for a given mass of gas:

$$\frac{pv}{T} = \text{const.}$$

Avogadro's law states that gram molecules of different gases at constant temperature and pressure occupy equal volumes (22.41 litres at a temperature of 0°C and a pressure of 1 atmosphere\*). Hence, for one gram mole, the constant in Clapeyron's equation is a universal constant. It is designated by the letter  $R$  and is

\* A physical atmosphere is meant here (1 physical atmosphere = 1.033 engineering atmospheres =  $1.01 \times 10^6$  N/m<sup>2</sup>).



called *the universal gas constant*. For a mole of any gas, the equation assumes the form

$$pv = RT.$$

Here  $v$  is the volume of a mole of gas. The constant  $R$  has the dimensions of work per mole degree. Expressed in various units, its value is

$$R = 8.31 \times 10^7 \frac{\text{ergs}}{\text{mole } K} = 8.31 \frac{\text{J}}{\text{mole } K} = 0.0821 \frac{\text{atm litres}}{\text{mole } K} = 2 \frac{\text{calories}}{\text{mole } K}.$$

Since the volume of an arbitrary mass of gas is  $V = \mu v$ , where  $\mu$  is the number of moles, the equation of state for a rarefied gas assumes the following form in the most general case:

$$pV = \mu RT \quad \text{or} \quad pV = \frac{m}{M} RT.$$

Here  $m$  is the mass and  $M$  is the molecular weight.

This equation yields the following convenient formula for the gas density  $\rho$ :

$$\rho = \frac{Mp}{RT}.$$

Gases obeying the equation of the gas state are called *ideal* gases. The simplicity of the equation would in itself be sufficient grounds for using this term. However, as we shall see below (p. 143), this equation may be derived by assuming the gas to be represented in an ideal system. An ideal gas is then a system of molecules whose dimensions and forces of attraction may be neglected.

For ideal gases, the coefficients of dilation, change of pressure and compressibility are given by the following simple formulas:

$$\alpha = \beta = \frac{1}{T}, \quad \kappa = \frac{1}{p}.$$

At a temperature of  $0^\circ\text{C}$  ( $T = 273.1 \text{ K}$ )

$$\alpha = \beta = \frac{1}{273.1} K^{-1} = 0.00366 K^{-1}.$$

The following data show to what extent certain actual substances approach the ideal condition:

	$\alpha$ at $V = \text{const}$
Hydrogen . . . . .	$3,660 \times 10^{-6}$
Helium . . . . .	$3,660 \times 10^{-6}$
Nitrogen . . . . .	$3,674 \times 10^{-6}$
Carbon dioxide . . . . .	$3,726 \times 10^{-6}$
Air . . . . .	$3,674 \times 10^{-6}$

Gaseous substances under a pressure that considerably exceeds atmospheric cease to obey the formulas of an ideal gas. Calculations may lead to errors of several per cent at pressures of only tens of atmospheres.

An important conclusion to be drawn from the study of rarefied gases is that, generally speaking, any of these gases—and not only hydrogen—may be used as a basis for determining temperature. Hydrogen provides no particular advantage over other rarefied gases. It can, therefore, be said that the temperature scale

adopted in physics is not a hydrogen scale, but rather a scale of pressures for an ideal gas. Herein lies the advantage of the method adopted for determining temperature, namely, the existence of a large class of substances leading to temperature scales that are identical. The kinetic molecular basis for the selection of the temperature scale will be given below (p. 144).

#### Sec. 57. THE EQUATIONS OF STATE OF ACTUAL GASES

The equation of the gas state begins to yield very rough results for gases at high pressures, steam close to saturation and a number of other cases. Other equations of state are, therefore, required for such cases. Some are determined experimentally, while others (the most well known being Van der Waals' equation) have, qualitatively, a theoretical basis. In any case, the validity of one or another equation can be established only through comparison of the results calculated by means of the equation with the results obtained experimentally. We shall now give some examples of equations of state.

Naturally, the simplest correction that can be introduced into the equation for ideal gases takes account of the volume of the gas molecules. It is evident that a gas cannot be compressed to zero volume even if the pressure is infinitely large. Hence, the equation of state may be written in the form

$$p(v - b) = RT,$$

where  $b$  is a constant that takes account of the finite volume of the molecules.

The greater the number of constants introduced into the equation of state the easier it is to achieve close agreement between experimental and calculated values, but the more difficult it is to predict changes by means of the formula. Excellent agreement with experiments over a broad interval of values for the parameters of state is obtained by means of the formula proposed by Beattie and Bridgeman. It contains five constants— $A$ ,  $B$ ,  $a$ ,  $b$  and  $c$ —descriptive of the substance:

$$p = \frac{RT(1-\epsilon)}{v^2} (v + B') - \frac{A'}{v^2},$$

where

$$A' = A \left(1 - \frac{a}{v}\right), \quad B' = B \left(1 - \frac{b}{v}\right), \quad \epsilon = \frac{c}{vT^3}.$$

Dieterich's formula contains three constants— $a$ ,  $b$  and  $s$ :

$$p(v - b) = RT e^{-\frac{a}{RT^2 v}}.$$

Van der Waals' equation contains two constants— $a$  and  $b$ :

$$\left(p + \frac{a}{v^2}\right)(v - b) = RT.$$

The merit of the last equation is that it correctly reflects the general character of the dependence between the parameters for all gaseous substances. However, for a given substance, it is not possible to select constant values for  $a$  and  $b$  in such a manner that the calculations agree closely with measurements over a broad interval.

Van der Waal's equation is based on the following: the pressure satisfies the equation of the gas state, i.e.,  $p = \frac{RT}{v}$ , when the forces of attraction between the molecules are neglected. But due to the mutual attraction of the molecules, the

pressure on the walls of a vessel should decrease by some value  $p'$ . Thus,  $p = \frac{RT}{v} - p'$ . Now, taking the finite volume of the molecules into consideration,

$$p = \frac{RT}{v-b} - p' \quad \text{or} \quad (p + p')(v - b) = RT.$$

Why does  $p' = \frac{a}{v^2}$ ? Here, we reason as follows: let us consider the gas volume divided into two parts. One part, then, attracts the other. The forces of attraction are proportional to the number of molecules in the left-hand part and to the number of molecules in the right-hand part. In other words, the forces of attraction are proportional to the square of the density, i.e., inversely proportional to the square of the volume.

The forces of attraction between molecules will be discussed in more detail in Part III.

# Thermodynamic Processes

## Sec. 58. GRAPHICAL REPRESENTATION

If two parameters of state are given for a body, the third may be calculated by means of the equation of state. Thus, when one parameter (e.g., pressure) is plotted along one axis and a second parameter (e.g., volume) is plotted along the other axis, the state of the body is uniquely described by a plotted point.

To be sure, it has been assumed in the graphical representation that the state of the body is in equilibrium. Only then will the values of the parameters of state be the same throughout the volume of the system and are we justified in speaking of the temperature, pressure, density, etc., of the body (or system) as a whole.

The question arises: what kind of process can be involved if equilibrium states are being considered? The answer consists in the following: In a process that proceeds sufficiently slowly, the values of the parameters of state throughout the volume are *equal*. Such a process may be considered to be a continuous succession of equilibrium states. It is *reversible* and may occur in either direction. A process consisting of a succession of equilibria may proceed from state 1 to state 2 and then from state 2 back through all the intermediate states to state 1, without producing any changes in the surrounding medium.

A reversible process is an idealised process. Every actual process is in one way or another irreversible, depending on how far away the intermediate states of the process are from equilibrium.

This becomes clear from the following reasoning. Every establishment of equilibrium is irreversible. Many simple and familiar examples can be cited: the cooling of a body by placing it in a cooler surrounding, the "dissipation" of a mechanical deformation, e.g., the restoration of a compressed spring to its undeformed position upon being released, the spontaneous intermixing of two gases, etc. Reversible processes cannot proceed of themselves. They cannot be single processes occurring in a closed system.

An actual process does not consist of a succession of equilibria. Inevitably, such phenomena occur as those enumerated above. Hence, when the process is made to proceed in the reverse direction, it will never pass through exactly the same states as during the forward direction. When a gas is rapidly compressed, the pressure of the gas in layers close to the piston will be higher than elsewhere. On the other hand, during the reverse process—expansion of the gas—the pressure near the piston will be lower than elsewhere.

Nevertheless, in spite of the fact that reversible processes are an idealisation, they are of great interest since in many cases the difference between actual and reversible processes is insignificant. It all depends on the *relaxation time*, i.e., the time it takes for equilibrium to be established. This time varies within very broad limits—from  $\sim 10^{-16}$  sec, the time it takes for the pressure to equalise in a homogeneous gas, to minutes, hours, or even weeks, for processes involving heterogeneous substances.

Let us assume that a gas is compressed and the entire process takes one second. The relaxation time is an insignificant fraction of a second. We may therefore consider the actual process as a succession of equilibrium states and draw the curve on a graph showing  $p$  and  $v$ , or on some similar diagram. The same holds true for

all other processes in which the relaxation time is small with respect to the duration of the process.

Fig. 75 shows several curves representing simple processes. The coordinates of the graph give the pressure and the volume. In engineering thermodynamics, other coordinates are used in addition to these, but we need not discuss them. The vertical line 1 in the figure represents a process at constant volume. If the point generating the curve is moving upwards, the pressure is increasing; if the reverse is true the pressure is falling. It is clear that a change in temperature occurs during this process that is not "seen" on the diagram. The horizontal line 2 represents a process at constant pressure (isobaric process). Moving from left to right signifies expansion, while the reverse corresponds to compression. The curve designated by the figure 3 corresponds to an expansion accompanied by a drop in pressure, while curve 4 represents an expansion in spite of the increasing pressure. The change of temperature in any process may be calculated by means of the equation of state.

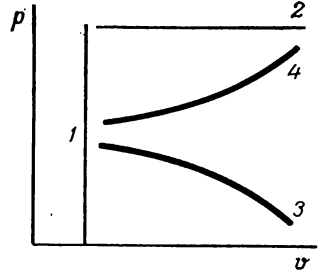


Fig. 75

In most thermodynamic processes, all the parameters of state are changing simultaneously. Nevertheless, a number of simple but at the same time practically important exceptions may be singled out. These include the processes mentioned above, namely, at constant volume (isochoric) and at constant pressure (isobaric), as well as the processes occurring without heat exchange (adiabatic) and at constant temperature (isothermal).

#### Sec. 59. WORK AND CYCLES

In mechanics, work is usually represented as the product of a force and a distance. In thermodynamics, we are usually interested in the work of changing the volume of a body. Fig. 76 shows the shape of a body in two states. The volume is shown to have changed from  $v_1$  to  $v_2$ . The total work of changing the volume may be considered as the summation of the work expended in displacing the elements of area  $dS$  by the distance  $dl$ . If the applied forces are perpendicular to the surface, the work of displacing element is equal to  $f dl$  or, in terms of pressure,  $p dS dl$ .

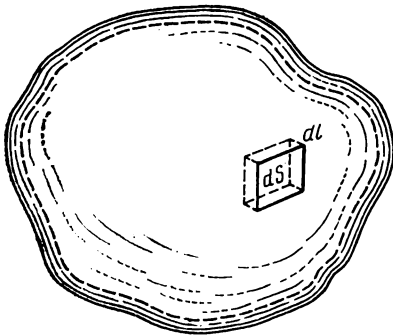


Fig. 76

Thus,

$$dA = p dv,$$

where  $dv$  is an element of volume change. It is evident that the total work is given by the definite integral:

$$A = \int_{v_1}^{v_2} p dv.$$

In a pressure-volume diagram, the work of compression or expansion can be represented geometrically. It is simply the area under the curve (bounded on the

left and the right by vertical lines through the points representing the initial and final values of the volume).

If the pressure during the process of compression or expansion remains unchanged and if, moreover, it is the same at all points on the surface, then  $p$  may be brought out from under the integral sign and the formula for work becomes

$$A = p (v_2 - v_1).$$

As we have already stated, work may be considered positive or negative depending on the convention adopted. We have assumed that work is positive when a body

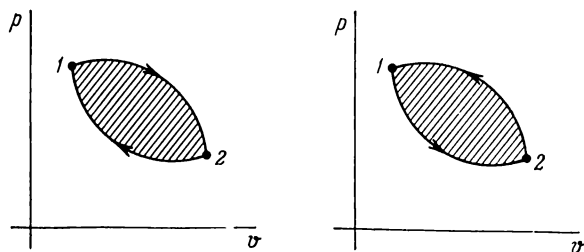


Fig. 77

does work on the surrounding medium, i.e., work of expansion. Accordingly, work of compression is negative.

If a body is transferred from state 1 to state 2 as a result of some process, and then transferred back to its original state via the same path, the total work of such a process is naturally equal to zero. Here, the work of expansion, performed on external bodies, is equal to the work of

compression, performed by external bodies on the system under consideration. However, the situation is completely different when the “forward” and “return” paths differ. Processes in which a body is returned to its original state via a different path are called *cyclical* processes. Fig. 77 shows two such cycles and the arrows indicate the direction of the processes. One process proceeds in a clockwise direction and the other in a counterclockwise direction. A process going from left to right signifies expansion. Thus, in the clockwise cycle the work of expansion is greater than the work of compression. In this case, work is performed on the surrounding medium. It is evident that in the counterclockwise cycle a certain amount of work is performed on the system under consideration. In either case, the work performed during a cycle is represented by the enclosed area (hatched in the figure).

#### Sec. 60. PROCESSES INVOLVING A CHANGE OF GAS STATE

Let us consider the relations for the four simplest processes involving a change of gas state, whereby, in the main, we shall restrict ourselves to gases obeying the equation of the gas state. It will be presently seen that knowing the equation of state for a substance and applying the first law of thermodynamics, a number of valuable conclusions may be drawn regarding the behaviour of the body under various conditions. The first law of thermodynamics for gases will be used in the form

$$\Delta Q = dU + p dv.$$

**The Isochoric Process.** At constant volume, the first law of thermodynamics assumes the form

$$\Delta Q = dU.$$

Heat exchange occurs between the system under consideration and the external medium, but no work is performed on the external medium or the system under consideration. Two possibilities exist: either (1) the body absorbs heat from the

surrounding medium and its internal energy increases or (2) the body gives up heat to the medium and its internal energy decreases.

The quantity of heat required to increase the temperature of a body by one degree at constant volume is called *the thermal capacity at constant volume* and is designated by the letter  $c$  with subscript  $v$ :

$$c_v = \left( \frac{dU}{dT} \right)_{v=\text{const.}}$$

If the dependence of the internal energy of the gas on the temperature is known, the thermal capacity  $c_v$  may be calculated.

At high temperatures, there is a linear dependence between the internal energy of a gas and the temperature, since the thermal capacity in this case does not depend on the temperature.

We are unable here to prove an important theorem. It follows, however, from the general laws of thermodynamics that if the dependence between  $p$  and  $T$  is linear, then  $c_v$  cannot depend on the volume. Since such a linear dependence exists for gases obeying the equation of the gas state and Van der Waals' equation, then  $c_v$  does not depend on  $v$  for gases and the phrase "at  $v = \text{const.}$ " may be omitted in the above formula. Thus,

$$c_v = \frac{dU}{dT} \text{ (for gases).}$$

If the dependence of  $c_v$  on temperature is only slight, the internal energy of a gas may be represented by the formula

$$U = c_v T + \text{const.}$$

For ideal gases, the constant does not depend on the volume and may be dropped.

For a gas obeying Van der Waals' equation, the constant equals  $-\frac{a}{v}$ . Thus,

$$U = c_v T \text{ for an ideal gas}$$

and

$$U = c_v T - \frac{a}{v} \text{ for a gas obeying Van der Waals' equation.}$$

We see that, in the case of an ideal gas, a change in the volume of the gas when the temperature is maintained constant does not involve a change in energy (see p. 130). If the molecules are drawn together with a force per unit area of  $p' = \frac{a}{v^2}$ , then upon expansion of the gas the energy increases by the amount of work done against this force, i.e., by

$$\int p' dv = -\frac{a}{v} + \text{const.}$$

**The Isobaric Process.** In this process, all three terms in the equation for the first law of thermodynamics are different from zero. The system exchanges heat and work with its surroundings without a change of pressure occurring in the system. This process usually involves absorption of heat by a body from its surroundings; however not all the heat serves to raise the internal energy of the body and, in part, it is returned to the surroundings in the form of mechanical work. We shall not consider other cases.

It is perfectly evident that the thermal capacity in this process will differ from that in the isochoric process considered above. In an isobaric process, the heat is used not only for raising the temperature. Hence,  $c_p$  (*thermal capacity at constant pressure*) must be greater than  $c_v$ . The difference may, in some cases, be calculated.

Let us divide both sides of the equation for the first law of thermodynamics by an incremental temperature:

$$c = \frac{\Delta Q}{dT} = \frac{dU}{dT} + p \frac{dv}{dT}.$$

This expression for the thermal capacity is valid for any process, including the isobaric process under consideration. For gases, this formula may be rewritten as follows:

$$c_p = c_v + p \frac{dv}{dT}.$$

For an ideal gas, the result obtained is very simple. Since  $pv = \mu RT$ , then  $\frac{dv}{dT} = \frac{\mu R}{p}$  and  $c_p = c_v + \mu R$ . Thus, the difference between the thermal capacities at constant pressure and at constant volume is equal to the number of moles of gas multiplied by the universal gas constant. Then, for molar thermal capacities,

$$c_p - c_v = R.$$

Since  $R \approx 2 \text{ cal/mole K} = 8.31 \text{ J/mole K}$ ,

$$c_p - c_v = 2 \text{ cal/mole K}.$$

**The Isothermal Process.** In order to avoid confusion, it should first be emphasised that constant temperature in no way signifies that no heat exchange occurs between the system and the surroundings. A system may absorb heat from the surroundings but not use it to raise the temperature. Thus, as is well known, the internal energy of a body may increase at constant temperature (e.g., melting ice). Moreover, for gas processes, there is another possibility (more important than the first): a system may return part of the heat received from the external surroundings in the form of mechanical work.

In the case of actual gases, both ways of expending the heat are entirely possible in an isothermal process. The heat transferred to a gas causes the gas to expand without the temperature being raised, whereupon (1) work is performed on the external surroundings and (2) the potential energy of the interacting molecules is increased.

In the case of an ideal gas, whose internal energy depends only on the temperature and therefore cannot change in an isothermal process, the first law of thermodynamics assumes a particularly simple form. Since  $dU = 0$ , then  $\Delta Q = \Delta A$ . Hence, either the system expands, absorbing heat from an external source and performing work on some object, or, conversely, the system contracts, releasing heat and obtaining energy in the form of mechanical work from external bodies. An ideal gas transforms energy in an isothermal process. It obtains energy from the surroundings in one form and returns all of it to the surroundings, but in another form.

It is not difficult in the case of an ideal gas to go over from the differential form  $\Delta Q = p dv$  to the integral form. The work (we can just as well say heat since work and heat are equivalent) of an isothermal expansion from volume  $v_1$  to volume  $v_2$  is

$$A = \int_{v_1}^{v_2} p dv.$$



Substituting for pressure from the equation of the gas state, and bringing the temperature out from under the integral sign since it is constant, we obtain

$$A = \mu RT \int_{v_1}^{v_2} \frac{dv}{v} = \mu RT \ln \frac{v_2}{v_1}.$$

It should be noted that the work of equal numbers of isothermal expansions at different temperatures differ, being greater the higher the temperature. Thus, doubling the volume of a mole of some ideal gas at a temperature of 300°K (room temperature) requires  $8.31 \times 300 \times \ln 2 = 1,730$  J of work, while at a temperature of 3,000°K the work required is ten times as much, i.e., 17,300 J.

An actual isothermal process may be difficult to achieve. In any case, in order for the process to be even approximately isothermal, the walls of the vessel through which the substance comes in contact with the surroundings must be perfectly heat-conducting. Moreover, the process must proceed very slowly, so that the heat (or work) is able to return to the surroundings in the form of work (or heat) instead of accumulating in the system.

**The Adiabatic Process.** Adiabatic compression and expansion occur in the absence of heat exchange with the surroundings. This may be achieved by providing conditions that are in a sense the reverse of those for an isothermal process, i.e., perfect thermal insulation must be provided and the process must proceed very rapidly, so that heat is not able to escape from the system or be transferred to the system. In the case of compression, in accordance with the first law of thermodynamics, which is now written in the form

$$p dv = -dU,$$

the mechanical work is converted into internal energy of the body. In the case of expansion, on the other hand, the work is performed at the expense of a decrease in the internal energy of the system under consideration.

For the three processes considered above, the changes occurring in the pressure, volume and temperature were quite apparent, and for gases followed directly from the equation of state. In an adiabatic process, the nature of the change in the parameters of state is not apparent, since all three parameters of state change. The simultaneous solution of two equations—the equations of the gas state and the first law of thermodynamics—enables us to determine the relationships. Since only the principle involved interests us, we shall restrict ourselves to an ideal gas in order to simplify the mathematical calculations. Using the expression for the thermal capacity of a gas at constant volume,  $\frac{dU}{dT} = c_v$ , and replacing the pressure  $p$  by  $\frac{\mu RT}{v}$ , we obtain:  $-\frac{\mu R}{c_v} \frac{dv}{v} = \frac{dT}{T}$ . Assume that in the initial state the gas parameters are  $v_1$ ,  $p_1$ ,  $T_1$  and in the final state  $v_2$ ,  $p_2$ ,  $T_2$ . Integrating the last equation from the initial to the final point of the adiabatic process, we obtain

$$-\frac{\mu R}{c_v} \int_1^2 \frac{dv}{v} = \int_1^2 \frac{dT}{T}; \quad -\frac{\mu R}{c_v} \ln \frac{v_2}{v_1} = \ln \frac{T_2}{T_1}.$$

Recalling that  $c_p - c_v = \mu R$  and introducing the designation  $\frac{c_p}{c_v} = \gamma$ , we obtain

$$\frac{T_2}{T_1} = \left( \frac{v_1}{v_2} \right)^{\gamma-1}.$$

It is seen from this equation that for adiabatic compression the temperature increases and for adiabatic expansion falls. Various examples can be cited. Thus, a gas is rapidly expanded when we desire to cool it and carbon dioxide gas escaping from a gas tank may turn into dry ice due to the tremendous drop in temperature of the expanding gas. On the other hand, adiabatic compression may be used, for example, to ignite some substance. The following experiment is often demonstrated: a wad of cotton wool soaked in ether is placed in a vessel containing air. Rapid compression of the air by means of a plunger causes the cotton wool to burst into flame.

Since we wish to represent the gas process on a pressure-volume diagram, it is necessary to convert the above equation of the adiabatic process into an appropriate form. Substituting for the temperature by means of the equation of the gas state, we obtain

$$p_1 v_1^\gamma = p_2 v_2^\gamma.$$

Comparing this equation with the Boyle-Mariotte law for an isothermal process, important differences may be observed in the nature of the pressure change for compression or expansion. For an isothermal expansion or compression, the product  $p v$  remains unchanged, while for an adiabatic process the product  $p v^\gamma$  remains unchanged. Since  $\gamma > 1$ , the adiabatic on the diagram is steeper than the isothermal.

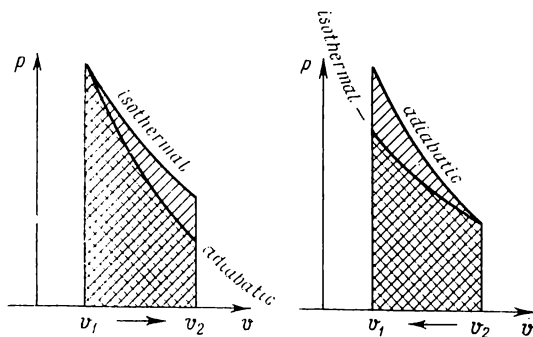


Fig. 78

When the volume is reduced to one-half in an isothermal process the pressure doubles, while in an adiabatic process the increase is greater. For example, in the case of most diatomic gases, for which  $\gamma = 1.4$ , when the volume is reduced to one-half, the pressure increases to 2.63 times its original value.

It has already been emphasised that both processes are of an

ideal character and that the requirements for the creation of the ideal conditions of these processes are opposite. Therefore, it is evident that gas processes under actual conditions yield intermediary curves between the adiabatic and isothermal curves.

The difference between the adiabatic and isothermal curves may be easily visualised as follows: for adiabatic compression, the gas becomes heated, so that for one and the same reduction in volume the increase in pressure is greater in the adiabatic process, since heating at constant volume leads to an increase in temperature.

Fig. 78 shows that the work of isothermal expansion is greater than the work of adiabatic expansion. On the other hand, the work of isothermal compression is less than the work of adiabatic compression. We are assuming, naturally, that the initial points of the processes coincide.

The work of an adiabatic process may be determined graphically or by means of formulas. From the first law of thermodynamics, it follows that in adiabatic processes the work must equal the change in internal energy:

$$A = \int_1^2 p dv = U_1 - U_2.$$

In the case of ideal gases, the difference in energy is calculated simply as  $U_1 - U_2 = c_v (T_1 - T_2)$ . Thus, for ideal gases, the work may be calculated by means of this formula.

Another method may be used to determine the work in an adiabatic process. Since the equation  $p_1 v_1^\gamma = p v^\gamma$  holds for any intermediate point of the process, where the symbols without subscripts designate the current values of pressure and volume respectively, the work integral may be written in the form

$$A = p_1 v_1^\gamma \int_{v_1}^{v_2} \frac{dv}{v^\gamma}.$$

Upon integrating from the initial to the final point of the process, we obtain

$$A = \frac{p_1 v_1^\gamma}{\gamma - 1} \left( \frac{1}{v_1^{\gamma-1}} - \frac{1}{v_2^{\gamma-1}} \right).$$

Naturally, this formula is equivalent to  $A = c_v (T_1 - T_2)$ , which is easily demonstrated by using the equation of state of an ideal gas and converting the above formula (taking  $\frac{1}{v_1^{\gamma-1}}$  out of the brackets) to the form

$$A = \frac{\mu R T_1}{\gamma - 1} \left[ 1 - \left( \frac{v_1}{v_2} \right)^{\gamma-1} \right].$$

Depending on the given data, one formula may be more conveniently used than the other.

Let us illustrate by a simple numerical example the statement made in Chapter IX to the effect that the increments  $\Delta Q$  and  $\Delta A$  (note:  $\Delta Q \sim \Delta A$ ) are not total differentials, i.e., they do not characterise the change of state of a system.

Assume state (1) of a mole of hydrogen (Fig. 79) is characterised by the following data:  $v_1 = 0.02 \text{ m}^3$ ,  $T_1 = 300 \text{ K}$  and  $p_1 = \frac{RT_1}{v_1} = 125,000 \text{ J/m}^3$  (here  $R = 8.31 \text{ J/K}$ .) Now,

$c_p = c_v = R$  and, since hydrogen is a diatomic gas,  $\frac{c_p}{c_v} = 1.4$ . Hence  $c_p = 29.4 \text{ J/K}$  and  $c_v = 21 \text{ J/K}$ .

We shall now consider three possible paths by means of which the gas can change to state (3), where  $v_3 = 0.04 \text{ m}^3$ ,  $T_3 = 300 \text{ K}$  and  $p_3 = 63,000 \text{ J/m}^3$ .

*Path 1-3.* The work along the isotherm  $A_{1-3} = RT_1 \ln \frac{v_3}{v_1} = 1,700 \text{ J}$ . These 1,700 J are absorbed from the hot body, and the internal energy  $U = \text{const}$  since  $T_1 = T_3$ .

*Path 1-2-3.* Here, (1-2) is an isobar. Hence,  $T_2 = 600 \text{ K}$ . The heat absorbed from the hot body is  $Q_{1-2} = c_p (T_2 - T_1) = 8,600 \text{ J}$  and the work against the external forces is  $A_{1-2} = p_1 (v_2 - v_1) = 2,300 \text{ J}$ . Therefore, the internal energy of the gas increases by  $\Delta U = 8,600 - 2,300 = 6,300 \text{ J}$ . Process 2-3 constitutes isochoric cooling and  $Q_{2-3} = c_v (T_2 - T_3) = 6,300 \text{ J}$  of heat are transmitted to the cold body. Since  $v_2 = v_3$ , no mechanical work is performed.

Thus, along the path 1-2-3, the hot body gave up 8,600 J, 2,300 J of work was performed and the cold body absorbed 6,300 J. Along the path 1-3, the hot body gave up 1,700 J, 1,700 J of work was performed and no change in the state of the cold body occurred. However, the change in the state of the gas was the same for both cases.

*Path 1-4-3.* Here (1-4) represents an adiabatic process, while (4-3) is an isobar. Now,  $\frac{v_4}{v_1} = \left( \frac{p_1}{p_4} \right)^{\frac{1}{\gamma}} = 2^{\frac{1}{\gamma}}$ , so that  $v_4 = 0.020 \times 2^{\frac{1}{\gamma}} \text{ m}^3$ . From the relation  $\frac{T_4}{T_1} = \left( \frac{v_1}{v_4} \right)^{\gamma-1}$ , we find

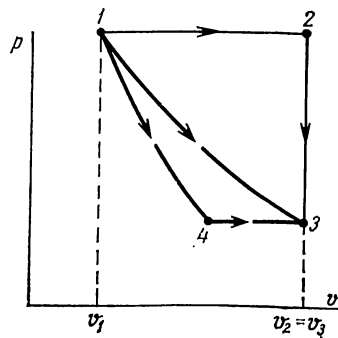


Fig. 79

that  $T_4 = 300 \times 2^{\frac{1}{\gamma}-1}$ . Along the path 1-4, the work against the external forces is performed only at the expense of a decrease in the internal energy:  $A_{1-4} = -c_v(T_4 - T_1) = 6,300(1 - 2^{\frac{1}{\gamma}-1})$  J. Along the path 4-3, the hot body gives up  $Q_{4-3} = c_p(T_3 - T_4) = 8,600(1 - 2^{\frac{1}{\gamma}-1})$  J of heat and the work against the external forces equals  $A_{4-3} = p(v_3 - v_4) = 2,300(1 - 2^{\frac{1}{\gamma}-1})$  J.

Therefore, along the path 4-3, the internal energy increases just by  $6,300(1 - 2^{\frac{1}{\gamma}-1})$  J. Thus the path 1-4-3 has also not led to a change in the internal energy of the gas, which is uniquely determined by the temperature.

**Measuring the Thermal Capacity of Gases.** It would appear that the easiest way to determine the thermal capacity of a gas is to fill a container with the gas to be measured and immerse it in a calorimeter. However, this does not take into account the fact that the thermal capacity of a gas is very small with respect to the thermal capacity of a container, no matter what solid material is used to make the container. Therefore, the thermal capacity of a gas is not measured at constant volume, but rather at constant pressure. For this purpose, gas under constant pressure moves in a coiled pipe that passes through the calorimeter. By means of a thermocouple, the temperature of the gas is measured at the input and output of the calorimeter. After preliminary heating, the gas enters the calorimeter and transfers part of its heat to the water. Knowing the quantity of gas passing through the container during a particular period of time, and the quantity of heat transferred to the water of the calorimeter during the same interval of time, the thermal capacity of the gas at constant pressure,  $c_p$ , can be easily determined. This is done by dividing the quantity of heat by the mass of flowing gas times the difference in gas temperature between input and output.

To determine the thermal capacity at constant volume, we use the ratio of thermal capacities, i.e., Poisson's coefficient:  $\gamma = \frac{c_p}{c_v}$ . Many methods of determining  $\gamma$  have been proposed, some of them being based on the measurement of the volume and pressure of the gas in a succession of states for an adiabatic process. Other relationships between the thermal capacities may also be used, e.g., the relationship defining the difference between the thermal capacities  $c_p$  and  $c_v$ .

The thermal capacities of various gases are given in the table.

Gas	$c_v$ J/K mole	$c_p$ J/K mole	$\gamma$
Helium He . . . . .	12.5	20.9	1.67
Hydrogen H <sub>2</sub> . . . . .	20.4	28.8	1.41
Nitrogen N <sub>2</sub> . . . . .	20.39	28.6	1.41
Oxygen O <sub>2</sub> . . . . .	20.9	28.9	1.40
Water vapour H <sub>2</sub> O . . . . .	27.8	36.2	1.31
Methane CH <sub>4</sub> . . . . .	27.3	35.6	1.30
Ethyl alcohol C <sub>2</sub> H <sub>5</sub> OH . . . . .	79.4	87.7	1.11

## Sec. 61. JOULE-THOMSON PROCESS

This is the process in which gas is allowed to flow through a small opening from a region of high pressure  $p_1$  into a region of low pressure  $p_2$ . The vessel in which the process takes place is thermally insulated from the surroundings.

In accordance with the conditions of the process,  $p_1$  and  $p_2$  must not change. This is done by having both pistons (Fig. 80) move to the right, corresponding to the passage of gas into the region of low pressure.

$M$ , the mass of gas that is moved from left to right, does not maintain constant volume, but changes from  $v_1$  to  $v_2$ , for it has entered a region of different pressure. This transition is accomplished under the action of the left piston and the counteraction of the right one. The left piston does work at the constant pressure  $p_1$ . This work is equal to  $p_1\Delta v$ , where  $\Delta v$  is the change in the volume of the gas to the left of the partition. But  $v_1$  is the change in volume on the left, so that the work done by the left piston is  $p_1v_1$ . This right piston does negative work, which in this case is equal to the product of the pressure  $p_2$  and the incremental volume  $v_2$ . Thus, when the mass of gas  $M$  is transferred from the left region to the right one, the work performed is  $p_1v_1 - p_2v_2$ . The law of conservation of energy requires that the internal energy of the gas change by this same amount. Therefore,

$$U_2 - U_1 = p_1v_1 - p_2v_2.$$

This formula is valid for any mass of gas. This means that in the process of moving a gas from one vessel into another the quantity

$$U + pv = \text{const}$$

(called the *heat function* or *enthalpy*) does not change.

For an ideal gas,  $U$  and  $pv$  both depend only on the temperature. Thus, during a Joule-Thomson process, the temperature of an ideal gas does not change.

For actual gases, the situation is different. If the gas is not ideal the temperature may increase or decrease during a Joule-Thomson process, depending on the nature of the interacting forces between the molecules.

It is noteworthy that a particular gas may behave differently at different temperatures. At a high temperature, the temperature increases during a Joule-Thomson process, while at a low temperature, it decreases. The *point of inversion* corresponds to the temperature at which the change in sign occurs. This temperature for oxygen and nitrogen is above room temperature. Therefore, the air is cooled during a Joule-Thomson process conducted at room temperature—not to speak of lower temperatures. For hydrogen, the inversion temperature is very low. The Joule-Thomson effect below the inversion temperature finds application industrially in the liquefaction of gases.

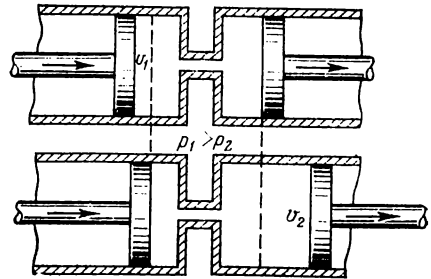


Fig. 80

# Entropy

## Sec. 62. THE PRINCIPLE OF ENTROPY EXISTENCE

In the middle of the last century, an important discovery was made regarding reversible thermodynamic processes. It was found that side by side with internal energy a body has yet another remarkable function of state, namely, entropy. Just as in the case of internal energy, entropy includes an arbitrary constant. Experiments yield the incremental difference in entropy. If a body or system absorbs the heat  $\Delta Q$  during an infinitesimally small transition from one state to another at a temperature  $T$ , the ratio  $\frac{\Delta Q}{T}$  is a total differential of some function  $S$ . This function is *the entropy* and is thus determined by the two following equivalent equations:

$$dS = \frac{\Delta Q}{T} \quad \text{and} \quad S_2 - S_1 = \int_1^2 \frac{\Delta Q}{T}.$$

The statement that a function exists whose differential is  $\frac{\Delta Q}{T}$  is known as *the principle of entropy existence*. It is one of the most important laws of nature and an essential part of the second law of thermodynamics, which will be discussed below. The discovery of this principle, as well as the entire second law of thermodynamics, is primarily associated with the names of Carnot and Clausius. In spite of its somewhat abstract nature, the essence of the principle is easily understood and may be summarised as follows: a body may change from one state to another in an infinite number of ways (represented on a diagram by the various curves connecting the same initial and final points); and, although the body may absorb various amounts of heat during such transitions, the integral  $\int_1^2 \frac{\Delta Q}{T}$  will in all cases have the same value.  $\frac{\Delta Q}{T}$ , the ratio of the quantity of heat to the temperature at which this heat was absorbed, is sometimes called *the reduced heat*. Since an integral may always be approximately represented as a summation, the change of entropy in transferring from one state to another is equal to the summation of the reduced heats. Let us assume that the body absorbs a calorie per degree as it is uniformly heated from 20°C to 25°C. The increase in entropy is then

$$S_1 - S_2 \approx \frac{1 \text{ J}}{293.5 \text{ K}} + \frac{1 \text{ J}}{294.5 \text{ K}} + \frac{1 \text{ J}}{295.5 \text{ K}} + \frac{1 \text{ J}}{296.5 \text{ K}} + \frac{1 \text{ J}}{297.5 \text{ K}}.$$

For isothermal processes, the expression for the change in entropy is very simple:

$$S_2 - S_1 = \frac{Q}{T},$$

where  $Q$  is the heat absorbed during the process. Thus, when 1 kg of ice melts the entropy of the substance increases by

$$\frac{33.6 \times 10^4 \text{ J}}{273 \text{ K}} = 1,230 \frac{\text{J}}{\text{K}}$$

In applying the concept of entropy, the value of entropy at any state (e.g., boiling water or melting ice) may be adopted as zero entropy. However, in some cases, the value of the entropy of a substance at absolute zero is adopted as zero entropy. There is, incidentally, some theoretical basis for this (Nernst's theorem), but we shall not go into it.

Assuming  $S = 0$  at  $T = 0$ , the entropy of a substance at the temperature  $T$  may be determined by the formula

$$S = \int_0^T \frac{c_p dt}{T},$$

if the heating occurs at constant pressure. As can be seen, the dependence of the thermal capacity on the temperature must be known in order to determine the entropy.

The entropy may be easily calculated (except for the arbitrary constant) if the equation of state of the substance is known. By definition,  $dS = \frac{\Delta Q}{T}$ . Substituting the value for  $\Delta Q$  obtained from the equation for the first law of thermodynamics, we obtain

$$dS = c_v \frac{dT}{T} + p \frac{dv}{T}.$$

By means of the equation of the gas state, we can eliminate the pressure from this equation, obtaining:  $dS = c_v \frac{dT}{T} + \mu R \frac{dv}{v}$ . Taking the indefinite integral, we obtain an expression for the entropy that includes an arbitrary constant:

$$S = c_v \ln T + \mu R \ln v + \text{const.}$$

It is also possible to take the definite integral of  $dS$ , where the limits are two states. We then obtain the following expression for the entropy difference between the two states:

$$S_2 - S_1 = c_v \ln \frac{T_2}{T_1} + \mu R \ln \frac{v_2}{v_1}.$$

This is the expression for the entropy of ideal gases. It is seen from the formula that the entropy increases as the temperature and the volume of the gas increase. Naturally, this is in agreement with the general statement that the entropy increases when heat is transferred to the body.

*Example:* Using the example on p. 129 (Fig. 79), we shall show that the entropy is indeed a function of the state of a system:

*Path 1-2-3.* The change of entropy

$$S_2 - S_1 = c_v \ln \frac{T_2}{T_1} + R \ln \frac{v_2}{v_1} = 20.74 \ln 2 + 8.38 \ln 2 = 29.36 \ln 2 \frac{\text{J}}{\text{K} \times \text{mole}}.$$

The change of entropy  $S_3 - S_2 = 20.74 \ln 2 = -20.74 \ln 2 \frac{\text{J}}{\text{K} \times \text{mole}}$ . The total change of entropy along path 1-2-3 is  $S_3 - S_1 = 8.38 \ln 2 \frac{\text{J}}{\text{K} \times \text{mole}}$ .

$$\text{Path 1-3. } S_3 - S_1 = 8.38 \ln 2 \frac{\text{J}}{\text{K} \times \text{mole}}.$$

*Path 1-4-3.* Since (1-4) is adiabatic,  $S_4 - S_1 = 0$ .

$$S_3 - S_4 = c_v \ln \frac{T_3}{T_4} + 8.38 \ln \frac{v_3}{v_4} = 20.74 \ln 2^{1-\frac{5}{7}} + 8.38 \ln 2^{1-\frac{5}{7}} = 29.36 \ln 2 \frac{\text{J}}{\text{K} \times \text{mole}}.$$

It is seen, indeed, that no matter how the transition of the gas from state (1) to state (3) is effected the change of entropy is the same.

## Sec. 63. THE PRINCIPLE OF INCREASING ENTROPY

As already stated, reversible processes, strictly speaking, do not exist. However, many processes occur that do not, practically, differ from reversible ones. But there are some processes that are always unidirectional and as a result can never be made reversible. Thus, gas may expand of itself, but it cannot be compressed without the application of an external force. Heat may spontaneously pass from a hot body to a colder one, but it can pass from a cold body to a hotter one only if work (e.g., electric energy) is expended. In the presence of friction, the kinetic energy of macroscopic motion is always converted into internal energy, but the reverse process never occurs spontaneously. All other irreversible processes are in the final analysis based on the fact that in each of them, to one degree or another, one of the enumerated unidirectional processes occurs. In actual processes, it is impossible to avoid spontaneous expansion, friction and thermal dissipation.

Do not all the enumerated unidirectional processes have a common characteristic? As a matter of fact they do: in all unidirectional processes, the entropy increases.

In the case of heat exchange between two bodies, the overall change in entropy of the entire system is  $S_2 - S_1 = \frac{Q_1}{T_1} + \frac{Q_2}{T_2}$ , where  $Q_1$  is the heat absorbed by the colder body and  $Q_2$  is the heat given up by the hotter body.

If  $T_2$  is greater than  $T_1$ , then  $Q_1 = -Q_2 > 0$ , since heat transferred to a body is considered positive. Hence,  $S_2 - S_1 = Q_1 \left( \frac{1}{T_1} - \frac{1}{T_2} \right) > 0$ , i.e., during heat exchange, there is an increase in the overall entropy of the system in which the heat exchange occurs.

Let us take another case. Assume intensive mechanical motion (e.g., rotation of a wheel) takes place in a vessel containing gas. The volume does not change, but the temperature increases and, hence, the entropy changes by  $S_2 - S_1 = c_v \ln \frac{T_2}{T_1}$ , i.e., it also increases.

Finally, upon expansion into an evacuated vessel at constant temperature, the increase in entropy,  $S_2 - S_1 = \mu R \ln \frac{v_2}{v_1}$ , is again positive. Thus, in all unidirectional processes, the entropy of the system increases.

It is easily seen that this conclusion regarding all irreversible processes is of great importance. Since each irreversible process is accompanied by unidirectional effects serving to increase the entropy, the increase in entropy in an irreversible process is greater than the increase that would have occurred if the process were reversible. Let  $\Delta Q$  be the heat absorbed by a body at temperature  $T$  in an irreversible process. If the process were reversible, the increase in entropy would equal  $\frac{\Delta Q}{T}$ . In an actual process, however, the increase in entropy is greater than this value:

$$dS > \frac{\Delta Q}{T}.$$

If the system is thermally insulated, then  $\Delta Q = 0$  and the above expression assumes the form  $dS > 0$ , i.e., in a thermally insulated system, only processes serving to increase the entropy are possible.

It is quite clear that entropy and internal energy are the most important functions determining a thermodynamic process. Thus, if entropy is analogous to the manager of a process, internal energy is analogous to the bookkeeper. While entropy determines the direction of flow of the process, the energy "meets the expenditures" of conducting it.



If in the above formula the symbol  $\gg$  is used instead of the symbol  $>$ , the law of entropy for reversible and irreversible processes may be described by the following simple formula:  $dS \gg \frac{\Delta Q}{T}$ . This formula expresses the essence of the *second law of thermodynamics*. For closed systems, the second law states that the entropy of a thermally isolated system increases or remains the same.

Both laws of thermodynamics may be combined in the single formula  $dS \gg \frac{dT + pdv}{T}$ , which is applicable to all practical thermodynamic problems.

The principle of entropy increase is applied to closed systems only. If a system associates with a medium, or, in other words, if we speak of an open system, then, of course, its entropy can decrease.

It will be shown below that the molecule ordering processes result in an entropy decrease. From a disordered system of small molecules received during the processes of nutrition and respiration, a living organism constructs highly organised structures, i.e., biological macromolecules (see Sec. 253). This leads to a decrease in organism entropy.

Consider a closed system: organism + medium, whose entropy must increase. It is obvious that entropy of the medium must increase so as to exceed the decrease in entropy of the organism.

The medium entropy increases at the expense of organism secretions.

If the process is steady, then

$$\left(\frac{dS}{dt}\right)_{\text{org}} = -\left(\frac{dS}{dt}\right)_{\text{med}}$$

We may state that the vital activity of the organism consists in passing an entropy flux of a substance through the organism. Here, the entropy of the substance entering the organism is less than that given off to the medium since the organism degrades food products.

*Examples.* 1. In the example on p. 50, we considered the nonelastic collision of a bullet with a ballistic pendulum and showed that, upon collision, 3,920 J of mechanical energy are dissipated in the bullet-pendulum system. This means that the bullet irreversibly transfers  $\Delta Q = 3,920$  J to the pendulum through heat conduction. If it is assumed that the process is isothermal (i.e., the thermal conductivity of the pendulum is extraordinarily high), and that the temperature is, say, 27°C, then the entropy of the system in this irreversible process will increase by

$$\Delta S = \frac{\Delta Q}{T} = 13.1 \text{ J/K.}$$

2. A rubber ball weighing 0.3 kg rises 1 metre off the floor after being dropped from a height of 2 metres. In this isothermal process (assume  $t = 27^\circ\text{C}$ ), we transfer  $\Delta Q = 2.96$  J irreversibly, i.e., the entropy of the ball-floor system increases by

$$\Delta S = 12.87 \times 10^{-3} \text{ J/K.}$$

If the ball and floor were absolutely elastic, the entropy would not have changed ( $\Delta S = 0$ ) and the motion of the ball would have continued eternally.

3. Let us consider the irreversible process involved in the transfer of heat from a steam boiler to a condenser. Assume that the steam boiler is at a temperature  $t_1 = 300^\circ\text{C}$  and the condenser at a temperature  $t_2 = 30^\circ\text{C}$ . For a boiler thermal capacity of 10,000 kW and an efficiency of 25 per cent,  $7.5 \times 10^6$  J will be transferred from the boiler to the condenser every second. Since the boiler loses heat, its  $\Delta Q$  will be negative, i.e., the boiler loses entropy. For the condenser, on the other hand, the entropy increases. However, since  $T_1 > T_2$ , the entropy of the boiler-condenser system will increase each second by

$$\Delta S = \Delta Q \left( \frac{1}{T_2} - \frac{1}{T_1} \right) = 11.8 \times 10^3 \text{ J/K.}$$

## Sec. 64. THE PRINCIPLE OF OPERATION OF A HEAT ENGINE

A heat engine converts heat into work. In other words, it takes heat from some bodies and transfers it to others in the form of mechanical work. In order to accomplish this conversion, we must have at our disposal two bodies at different temperatures, between which heat exchange is possible. The hotter body will be designated as the hot body and the colder one as the cold body. In the presence of two such bodies, the process of conversion of heat into work may be described as follows: a substance capable of expanding (the working substance) is brought into contact with the hot body. Heat  $Q_1$  is taken from the hot body and is expended on the work of expansion,  $A_1$ , which is transmitted to surrounding bodies. The working substance is then brought into contact with the cold body and transfers heat  $Q_2$  to it at the expense of the work  $A_2$  performed on the working substance by the external forces.

To obtain a continuously operating heat engine, the compression process must be concluded where the expansion process began. In other words, the overall process must be cyclic. The working substance returns to its initial state at the end of each cycle. Hence, the law of conservation of energy requires that the energy obtained from the surrounding bodies equal the energy transferred to the surrounding bodies. The working substance obtained the heat  $Q_1$  during expansion and the work  $A_2$  during compression. On the other hand, it gave up the work  $A_1$  during expansion and the heat  $Q_2$  during compression. Hence,  $Q_1 + A_2 = Q_2 + A_1$  or  $A_1 - A_2 = Q_1 - Q_2$ . When the cyclic process is conducted clockwise, the work of compression is less than the work of expansion. Therefore, the last equation expresses the simple fact that the network transmitted to a working substance by an external medium is equal to the difference in the heat absorbed from a hot body and given up to a cold body. Accordingly, the efficiency of the cycle and, hence, of the engine as a whole is

$$\eta = 1 - \frac{Q_2}{Q_1}.$$

The described process for the operation of a heat engine is, naturally, an abstract scheme. However, the essential features of every heat engine are incorporated in this scheme. An expanding and contracting gas or steam is the working substance, the surrounding medium plays the role of cold body, and a steam boiler, or a fuel mixture in internal combustion engines, serves as the hot body.

A refrigerating engine, in which the cycle is reversed, requires the same three system components. The principle of operation of this engine consists in the following: expansion of the working substance occurs when it is in contact with the cold body. Thus, the cold body is cooled even further, which is precisely the task of the refrigerating engine. Now, in order to complete the cycle, the working substance must be compressed and the heat given up by the cold body rejected. This is accomplished when the working substance is in contact with the hot body. Thus, the hot body becomes even hotter. The "unnatural" transfer of heat from a cooler body to a hotter body is at the "expense" of work. We see, then, that when the cycle is conducted counterclockwise, the relationship between the energy transferred to a medium and the energy absorbed from a medium, i.e.,  $Q_1 + A_2 = Q_2 + A_1$  or  $Q_2 - Q_1 = -(A_1 - A_2)$ , where as before the subscript 1 refers to the portion of the process occurring when in contact with the hotter body, has the following meaning: The quantity of heat removed from a system must be compensated for by an equal quantity of mechanical work.

The second law of thermodynamics imposes certain conditions on the operation of a heat engine. If a process is assumed to be reversible, the change in entropy of the working substance for the entire cycle should equal zero. Stated otherwise, the change in entropy for the expansion process must equal (except for reversed sign) the change in entropy for the contraction process, i.e.,

$$\int \frac{dQ}{T_1} = - \int \frac{dQ}{T_2}.$$

In the case of an irreversible process, the entropy of the closed system, consisting of the hot body, the cold body and the working substance, increases and, therefore,

$$\int \frac{dQ}{T_1} + \int \frac{dQ}{T_2} > 0.$$

(It should be recalled that  $Q$  is an algebraic quantity. Thus, heat entering the system is considered positive.) By evaluating these integrals for specific processes, it is rather simple in a number of cases to determine the maximum efficiency of one or another heat engine cycle.

#### Sec. 65. EFFICIENCY OF A CARNOT CYCLE

We shall now derive the expression for the efficiency of an ideal heat engine operating without losses in a reversible cycle.

Let us first consider the theoretical four-stroke *Carnot cycle* represented in Fig. 81. The Carnot cycle consists of two isothermals (for temperatures  $T_1$  and  $T_2$ ) and two adiabatics. Assume that the first stroke of the cycle is an isothermal expansion from state 1 to state 2—the working substance is in contact with the hot body whose temperature is  $T_1$  and the process takes place very slowly. When state 2 is reached, contact with the hot body is broken, the working substance is thermally isolated and it has the possibility of expanding further. Work occurs at the expense of the internal energy and the temperature of the working substance is allowed to drop to  $T_2$ . From this point (state 3), two-stroke contraction begins. The working substance comes in contact with the cold body at temperature  $T_2$  and isothermally contracts to state 4. Here, the working substance is again thermally isolated and the contraction continues, now adiabatically, with the working substance being heated, at the expense of performed work, to the initial temperature  $T_1$ .

The adiabatic processes in a Carnot cycle are of an auxiliary nature, enabling us to transfer from one isothermal to another. These processes do not enter into the energy balance, since  $c_v (T_1 - T_2)$ , the work of adiabatic expansion, and  $c_v (T_2 - T_1)$ , the work of compression, cancel each other.

In an adiabatic process, the entropy of a system does not change. During isothermal expansion, the entropy of the hot body decreases by  $\frac{Q_1}{T_1}$  and the entropy of the cold body increases by  $\frac{T_2}{Q_2}$ . The working substance returns to its initial

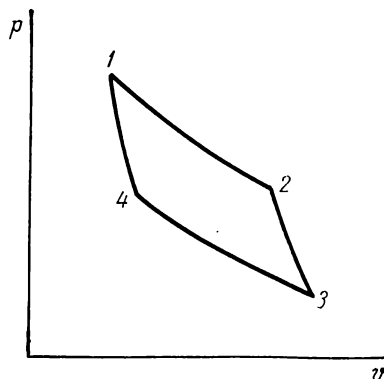


Fig. 81

state with its entropy unchanged. If the process is reversible, then  $\frac{Q_1}{T_1} = \frac{Q_2}{T_2}$ . For irreversible processes, the entropy of the entire system, consisting of the cold body, the hot body and the working substance, increases, i.e., the entropy increment  $\frac{Q_2}{T_2}$  is greater than the decrement  $\frac{Q_1}{T_1}$ :

$$\frac{|Q_2|}{T_2} > \frac{|Q_1|}{T_1}.$$

Thus,  $\left| \frac{Q_2}{Q_1} \right| \geq \frac{T_2}{T_1}$  and, therefore, the efficiency of a Carnot cycle is

$$\eta_{\max} = 1 - \frac{T_2}{T_1}.$$

The efficiency of the cycle is determined by the temperatures of the cold and hot bodies, respectively. The greater the drop in temperature the greater the efficiency of the engine. It is not difficult to see that the efficiency of a Carnot cycle is the maximum efficiency possible. There is no cycle better than the Carnot cycle and, in this sense, it serves as a model for designers of heat engines. They strive to make actual cycles approach the cycle of this ideal engine.

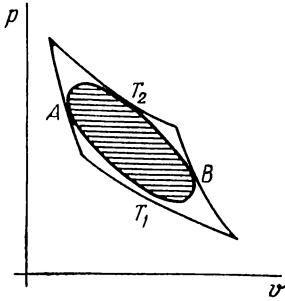


Fig. 82

It is not difficult to prove that the efficiency of a Carnot cycle is the optimum. Fig. 82 shows an arbitrary cycle inscribed in the Carnot cycle. The decrease in the entropy of the hot body may be represented by the integral

$$\int_A^B \frac{dQ}{T}$$

for which the inequality

$$\int_A^B \frac{dQ}{T} > \frac{1}{T_1} \int_A^B dQ = \frac{Q_1}{T_1}$$

is undoubtedly valid, since  $T_1$  is the largest value assumed by  $T$  in the integration. The increase in the entropy of the cold body is expressed by the integral

$$\int_B^A \frac{dQ}{T},$$

for which the inequality

$$\int_B^A \frac{dQ}{T} < \frac{1}{T_2} \int_B^A dQ = \frac{Q_2}{T_2}$$

is valid, since  $T_2$  is the smallest value assumed by  $T$  in the integration. For a reversible process,

$$\left| \int_B^A \frac{dQ}{T} \right| = \left| \int_A^B \frac{dQ}{T} \right|.$$

Therefore,

$$\frac{|Q_2|}{T_2} > \frac{|Q_1|}{T_1},$$

which yields the condition:

$$\eta_{\max} = 1 - \frac{T_2}{T_1}.$$

Thus, the Carnot cycle has the maximum efficiency of all possible cycles.

This maximum efficiency formula shows why steam engines have low efficiency. At  $T_2 = 300$  K and  $T_1 = 400$  K, the efficiency is 25 per cent. Moreover, this is the maximum efficiency, attained by an ideal reversible engine operating without any losses in energy. It is, therefore, not surprising that in actual steam engines the efficiency is below 10 per cent. Courses in steam engineering discuss means used to increase the efficiency. Clearly, the most important method is to increase the temperature of the hot body, i.e., the steam or fuel mixture.

## Sec. 66. THE SECOND LAW OF THERMODYNAMICS

As indicated above, the second law of thermodynamics states that the entropy in a thermally isolated system increases. This statement may appear to be somewhat abstract, but it is not the form in which this idea was expressed historically. In view of the tremendous importance of this law of nature, we shall briefly discuss other important formulations of the second law of thermodynamics and show that they are equivalent to the above.

Historically, the second law of thermodynamics was expressed in the form of Thomson's postulate on the impossibility of creating a perpetual engine of the second kind. A perpetual engine of the first kind creates work "out of nothing", i.e., its work violates the first law of thermodynamics. A perpetual engine of the second kind produces work by means of a periodically operating engine merely by absorbing heat from the surrounding medium. If such an engine were possible, it would be practically eternal, for the supply of energy in the surrounding medium is almost limitless and the cooling, say, of the oceans' waters by one degree would yield an inconceivably large amount of energy. The mass of water on the Earth is of the order of  $\sim 10^{18}$  tons. If this entire mass of water were cooled by only  $1^\circ$ , the heat released would be about  $10^{21}$  kcal  $= 4.18 \times 10^{24}$  J of heat, which is equivalent to the complete combustion of  $10^{14}$  tons of coal. Rolling-stock loaded with this quantity of coal would extend for a distance of  $\sim 10^{10}$  km, which is the order of magnitude of the dimensions of the solar system!

A perpetual engine of the second kind is a heat engine working with a hot body, but without a cold body. If such an engine were possible, it could work on a single stroke. A gas contained in a cylinder with a piston could indeed expand, but the operation of the engine would end there, since for the engine to continue operating, the heat absorbed by the gas must be transferred to a cold body. Formally, the formula for maximum efficiency shows that a perpetual engine of the second kind is impossible. In the absence of a temperature drop ( $T_2 = T_1$ ), the maximum efficiency is equal to zero.

It is impossible to design a periodically operating perpetual engine by combining an isothermal expansion with an adiabatic compression process. Such a process would not be possible even if we could make it reversible. For isothermal expansion of the working substance, the entropy decreases. Hence, the compression process would have to yield an increase in entropy. This, however, is not possible for an adiabatic process, since it proceeds at constant entropy.

The postulate of Clausius also completely corresponds to the formulation adopted here for the second law of thermodynamics. It states that heat cannot be transferred from a colder body to a hotter body without compensation. A process contradicting the postulate of Clausius would take place with a decrease in entropy. At the very beginning of our discussion, this was shown to be impossible.

We shall again return to the second law of thermodynamics in Sec. 77, where it will be discussed from the standpoint of the kinetic molecular theory.

# Kinetic Theory of Gases

## Sec. 67. GENERAL

If the molecules of a solid body are assumed to be contiguous, we can accurately determine their dimensions by X-ray analysis. Then by comparing these dimensions with the space available to a molecule in a gas, the fundamental properties of the gaseous state of matter may be immediately determined.

The largest linear dimensions of a diatomic molecule of oxygen is about  $4 \text{ \AA}$ . Nitrogen molecules have approximately the same dimension, but molecules of hydrogen are considerably smaller. The volume of an oxygen molecule is about  $10^{-23} \text{ cm}^3$ . Since under normal conditions there are  $2.7 \times 10^{19}$  molecules in  $1 \text{ cm}^3$  of oxygen, the space available to a molecule is about  $0.4 \times 10^{-19} \text{ cm}^3$ . Comparison of the volume of a molecule with the space available to it shows how little of the space is occupied by molecules. It is evident that for such a low density collisions between molecules will be relatively rare. On the average, the length of the path traversed by a molecule between consecutive collisions is  $1,000 \text{ \AA}$ . However, the velocity of a molecule is large, about  $500 \text{ m/sec}$ , so that on the average a collision occurs every ten-thousand millionth ( $10^{-10}$ ) of a second. It will be shown below how these figures were obtained.

Molecules begin to draw together only when the distances between them become comparable to their own dimensions. Therefore, for a large part of their path, molecules move rectilinearly and uniformly. Only when one molecule comes within range of another does the force of interaction become effective. Since the interaction occurs over an insignificantly small portion of the path, we can speak of a collision between the molecules. The interval of time during which molecules perceptibly interact—in other words, the impact time—is equal to about  $10^{-13} \text{ sec}$ . Thus, a molecule spends by far the greatest part of its “life” in free motion subject to inertia.

This is the situation for gases under normal conditions. An increase in pressure, leading to an increase in density may considerably alter the picture.

The internal energy of gases in which interaction between molecules occurs only for the time of instantaneous collision does not contain potential energy of interaction between molecules. Such gases are called *ideal*. The use of one and the same term a second time will be shown to be justified by demonstrating the validity of the equation of the gas state for such gases.

Thus, a gaseous substance consists of a tremendous number of minute particles that pass through large spaces without colliding, then collide like billiard balls and fly apart in different directions, with different velocities, until the next collision. If we were to trace the path of a single gas molecule (naturally, this can be done only mentally), we would find it moving now to the left or to the right, now forward or backward. Sometimes it would be moving with a large velocity and at other times it would be moving slowly. In view of the chaotic nature of thermal motion in a gas, the molecules of a free gas in thermal equilibrium may be considered to have uniform density distribution throughout its volume. Furthermore, at a given instant, there will undoubtedly be equal quantities of molecules moving in all directions. Other random events will similarly be uniformly distributed. For example, at all locations, equal numbers of molecules per second of observation will be travelling without collision a distance of  $100 \text{ \AA}$  to  $200 \text{ \AA}$ .

It must be realised, however, that these statements are of a statistical character. They are valid on the average, whereby the greater the number of gas molecules involved the greater the validity.

We assert, for example, that the number of molecules moving "to the right" is the same as the number moving "to the left". Naturally, this does not mean that the numbers are equal to within several units. The number of molecules involved is so large that not only is a difference of several units insignificant, but even a difference of several million is negligible percentagewise.

If numerous measurements are taken of the gas density of a given volume, the values obtained for the number of molecules will differ somewhat from measurement to measurement. From these data, we can determine the average value for the number of molecules in the volume under consideration. If it were possible to measure within an accuracy of even several thousand molecules, the individual measurements would oscillate, percentagewise, to an insignificant extent about this average value.

When it is stated that a number of molecules have such and such a velocity, or move in some direction or other, or collide in accordance with some mechanism or other, then the average value of the number is always meant. If the number of gas molecules is large, the deviations of the instantaneous values from the average, i.e., the *fluctuations*, are negligible. In a very rarefied gas, however, the fluctuations may be considerable.

It is shown in the theory of probability that, using absolute values, the average relative deviation of the gas density from the average is approximately equal to  $\frac{1}{\sqrt{n}}$ , where  $n$  is the number of molecules in a unit volume. Since there are  $2.7 \times 10^{19}$  molecules in 1 cm<sup>3</sup> of gas, the fluctuation of the gas density within one cubic centimetre amounts to

$$\frac{1}{\sqrt{2.7 \times 10^{19}}},$$

i.e.,  $2 \times 10^{-10}$  from the average value. It is evident that such deviations are beyond experimental observation.

This is how matters stand with respect to all gas properties that are determined by the average number of molecules.

The origin of the kinetic theory of gases dates back to Daniel Bernoulli (1700-1788). M. V. Lomonosov (1711-1765) also made substantial contributions to its development. In the 19th century, the kinetic theory of gases developed under Clausius (1822-1888), Maxwell (1831-1879) and Ludwig Boltzmann (1844-1906) and assumed its modern form.

## Sec. 68. MEAN FREE PATH

The distance traversed by a molecule between two consecutive collisions (the range of a molecule) is, naturally, a random quantity that may sometimes be very small or very large for individual molecules. However, in view of the chaotic nature of the particle motion, the average value of this quantity for a given gas state is undoubtedly constant. The mean free path or, for brevity, the range  $l$  is related to the average velocity  $v$  of the molecular motion and the average time  $\tau$  between two collisions by the simple relation:  $l = v\tau$ \*. Typical values for these quantities were cited on p. 140.

\* Since we are merely concerned with the determination of the connection between the physical quantities and not with the determination of the exact formulas, we shall not differentiate between average and root-mean-square velocities (see below).

The range of a molecule depends, in the first place, on the number of molecules in a unit volume of gas. Moreover, it is evident that the larger the dimensions of the molecule the smaller the mean free path.

In order to visualise the character of this relationship, let us consider a cylindrical volume of gas through which a molecule moves along the cylinder axis. What is the path taken by the molecule?

Molecules are not points. They have dimensions determined by the distances for which molecular interaction becomes effective.

On the basis of crystallochemical measurements (see p. 470), we may, with considerable accuracy, ascribe a certain form to molecules. At distances extending beyond the limits of the molecule's "boundary", the forces of interaction are, practically, not effective.

Let us project the maximum cross-section of the molecules on to the base of the cylinder. Each molecule will be projected differently. Since there are many molecules, the average cross-sectional area will characterise a molecule with sufficient accuracy. This average area of cross-section  $\sigma$  is called *the effective cross-section*.

A collision will surely occur along the length of the cylinder if the area of the cylinder base is completely filled with the cross-sections of the molecules. If the cylinder base is equal to  $1 \text{ cm}^2$ , cylinder length equal to  $l$ , and the number of molecules per unit volume equal to  $n$ , then there will be a total of  $nl$  molecules in the cylinder. The projections of the cross-sections of these molecules will completely cover the cylinder base when  $nl\sigma = 1$ . Under these conditions, the value of  $l$  will have an order of magnitude that is close to the average range of the molecule, i.e.,  $l \approx \frac{1}{n\sigma}$ . More rigorous calculations confirm the validity of this rough estimation. In the exact formula, the factor  $\sqrt{2}$  enters in the denominator:

$$l = \frac{1}{\sqrt{2} n \sigma},$$

where  $\sigma$  has a constant magnitude for a given gas. Thus, the mean free path is determined only by the density. A decrease in density by a factor of 100, for example, results in an increase in the mean free path by the same factor.

For air under normal conditions, the effective cross-section  $\sigma$  is approximately equal to  $5 \times 10^{-15} \text{ cm}^2$ . This is in excellent agreement with the dimensions of oxygen and nitrogen molecules obtained from crystal measurements. The maximum dimension is equal to  $4.3 \text{ \AA}$  and the minimum is a little less, namely,  $3 \text{ \AA}$ . The radius of a circle having an area of  $5 \times 10^{-15} \text{ cm}^2$  is  $4 \text{ \AA}$ .

We can determine the dimensions of molecules by studying crystals. However, the investigation of particle collisions may be viewed as a method of establishing the effective cross-section of particles. This method is of value in studying atomic nuclei (p. 432).

The mean free path under normal conditions is: in air— $600 \text{ \AA}$ , in nitrogen— $600 \text{ \AA}$ , in hydrogen— $1,100 \text{ \AA}$  and in helium— $1,800 \text{ \AA}$ .

## Sec. 69. GAS PRESSURE. ROOT-MEAN-SQUARE VELOCITY OF MOLECULES

Let us consider the problem of using the simplified concepts regarding the motion and interaction of gas molecules to express the gas pressure in terms of the quantities characterising the molecule.



Assume that we have a gas enclosed in a spherical tank of radius  $R$  and volume  $v$ . Disregarding collisions between gas molecules, we may adopt the following simple scheme for the motion of each molecule: a molecule moves rectilinearly and uniformly with some velocity  $v$ , strikes the wall of the vessel and rebounds at an angle equal to the angle of incidence (Fig. 83). Traversing chords of equal length,  $2R \sin \theta$ , time after time, the molecule strikes the wall of the vessel  $\frac{v}{2R \sin \theta}$  times per second. For each impact, the momentum of the molecule changes by  $2mv \sin \theta$  (see p. 50). The change in momentum per second is equal to  $\frac{mv^2}{R}$ .

We see that the angle of incidence cancels out. If the molecule strikes the wall at an acute angle, the impacts will occur often, but will be weak. If the angle of incidence is close to  $90^\circ$ , the molecule will strike the wall less often, but will make up for it by stronger impacts.

The change in momentum for each impact of the molecule on the wall contributes to the overall force of the gas pressure. It may be assumed in accordance with the fundamental law of mechanics that the force of the pressure is simply the change occurring in the momentum of all the molecules in one second:  $\frac{mv_1^2}{R} + \frac{mv_2^2}{R} + \dots$  or, factoring out the constants,

$$\frac{m}{R} (v_1^2 + v_2^2 + \dots).$$

Assuming  $n$  molecules are contained in the gas, we may introduce the concept of the average of the velocity squared of a molecule, which is determined by the formula

$$\bar{v^2} = \frac{1}{n} (v_1^2 + v_2^2 + \dots).$$

The expression for the force of the pressure may now briefly be written as follows:

$$F = \frac{mn\bar{v^2}}{R}.$$

Dividing this expression by  $4\pi R^2$ , the surface area of a sphere, we obtain the gas pressure:

$$p = \frac{nm\bar{v^2}}{4\pi R^3}.$$

Replacing  $4\pi R^3$  by  $3V$ , the following interesting formula is obtained:

$$pV = \frac{1}{3} nm\bar{v^2} \quad \text{or} \quad pV = \frac{2}{3} n \left( \frac{m\bar{v^2}}{2} \right).$$

Thus, the gas pressure is proportional to the number of gas molecules and to the average value of the kinetic energy associated with the translatory motion of a gas molecule.

A very important conclusion may be drawn by comparing the obtained equation with the equation of the gas state. Comparison of the right-hand members of the

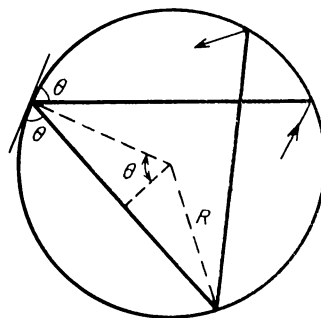


Fig. 83

equations shows that

$$\mu RT = \frac{2}{3} n \left( \frac{m\bar{v}^2}{2} \right) \quad \text{or} \quad \frac{m\bar{v}^2}{2} = \frac{3}{2} \frac{\mu R}{n} T,$$

i.e., the average kinetic energy of molecular translation depends only on the absolute temperature and, moreover, is directly proportional to it.

This conclusion shows that gases obeying the equation of the gas state are ideal in the sense that they approximate the ideal model of a group of particles whose interaction is insignificant. Further, it shows that the concept of absolute temperature, introduced empirically as a quantity proportional to the pressure of a rarefied gas, has a simple kinetic-molecular interpretation. The absolute temperature is proportional to the kinetic energy of molecular translation. The ratio  $\frac{n}{\mu} = N$  is known as Avogadro's number. It is the number of molecules in one gram molecule and is a universal constant:  $N = 6.02 \times 10^{23}$ . The reciprocal quantity  $\frac{1}{N}$  is equal to the mass of a hydrogen atom:

$$m_H = \frac{1}{N} = 1.66 \times 10^{-24} \text{ g.}$$

Another universal constant is the quantity

$$k = \frac{\mu R}{n} = \frac{R}{N} = 1.38 \times 10^{-16} \text{ erg/K,}$$

which is called *Boltzmann's constant*. Thus

$$\frac{m\bar{v}^2}{2} = \frac{3}{2} kT.$$

If the velocity squared,  $v^2$ , is represented by the sum of the squares of its components,  $v^2 = v_x^2 + v_y^2 + v_z^2$ , it is evident that the average energy for each component is

$$\frac{1}{2} kT.$$

This quantity may be described as the energy associated with one degree of freedom.

The universal gas constant is accurately known from experiments with gases. The determination of Avogadro's number and Boltzmann's constant, which are expressed in terms of each other, is a relatively difficult task involving delicate measurements.

These results put at our disposal useful formulas for calculating the average molecular velocity and the number of molecules in a unit volume.

Thus, for the average of the velocity squared, we obtain

$$\bar{v}^2 = \frac{3RT}{mN} = \frac{3RT}{M},$$

where  $M$  is the molecular weight. The square root of the average of the velocity squared is called *the root-mean-square velocity*:

$$v_{rms} = \sqrt{\frac{3kT}{m}} \quad \text{or} \quad v_{rm} = \sqrt{\frac{3RT}{M}},$$

i.e., the r-m-s velocity is directly proportional to the square root of the temperature and inversely proportional to the square root of the molecular weight. It is

easily determined that at room temperature oxygen molecules have a velocity of 480 m/sec and hydrogen molecules—1,900 m/sec. At the temperature of liquid helium, these molecules would have, respectively, velocities of 40 m/sec and 160 m/sec, while at the temperature of the surface of the Sun, namely 6,000°, these velocities would be 2,160 m/sec and 8,640 m/sec, respectively. These examples are unrealistic, however, for, at the temperature of liquid helium, oxygen and hydrogen solidify and no translatory motion of the molecules will occur, while at the temperature of the surface of the Sun the molecules disassociate into atoms.

We obtain the following simple expression for the number of molecules in a unit volume:

$$\frac{n}{V} = \frac{3p}{mv^2} = \frac{p}{kT}.$$

*Avogadro's law* follows from this and may be stated as follows: For equal pressures and temperatures, all gases contain the same number of molecules per unit volume. Thus, under normal conditions (at a pressure of 1 atm and a temperature of 0°C),  $2.683 \times 10^{19}$  molecules (*Loschmidt's number*) are contained in 1 cm<sup>3</sup>.

#### Sec. 70. INTERNAL ENERGY OF A GAS

The properties of monatomic gases are determined by the kinetic energy of translation of the molecules. An atom's internal energy does not affect the thermodynamics of the gas. Evidently, the internal energy need be considered only when the temperature of the gas is very high and collisions between atoms may lead to their excitation and ionisation. These processes will be discussed in detail later on.

Thus, the following formula for the internal energy of a monatomic gas will have very broad application:

$$U = N \frac{mv^2}{2},$$

where  $N$  is the number of molecules. Using the formulas of the previous article, we obtain for 1 mole of an ideal monatomic gas the expression

$$U = \frac{3}{2} RT.$$

Hence, for the thermal capacity of 1 mole of a monatomic gas, we obtain by means of the formulas of Sec. 60:

$$c_v = \frac{3}{2} R$$

and

$$c_p = \frac{5}{2} R.$$

The direct proportionality between the temperature and the internal energy, and hence the constancy of the thermal capacities of a monatomic gas, are valid for quite a broad interval of external conditions.

For polyatomic gases, such a simple picture is valid for a significantly narrower interval of temperatures, if valid at all. The reason for this is that the energy of a polyatomic molecule consists of the energy of translation, the energy of rotation and the energy of vibration of the molecule's components (i.e., the molecule's atoms) with respect to each other. Calculation of the average energy per molecule

becomes quite difficult. It turns out that the energy of a molecule is no longer linearly dependent on the temperature and, hence, that the thermal capacities are no longer constants independent of the magnitude of the temperature. Nevertheless, it is usually possible to find a narrow interval of temperatures in which the thermal capacities do not depend on the temperature. This occurs for such values of temperature at which the average energy of the molecule is not yet sufficient for the collisions of the molecule to lead to a change in its vibratory state. At the same time, this energy is sufficiently large so that the discrete (quantum) character of the energy of rotation is not felt. Jumping ahead and referring the reader to Fig. 277 (p. 479), it may be stated that linear dependence between energy and temperature, and constancy of thermal capacity, will occur when the quantity  $kT$ , descriptive of the order of magnitude of the translational energy of the molecule, is considerably greater than the distances between rotational energy levels and less than the distances between vibrational energy levels.

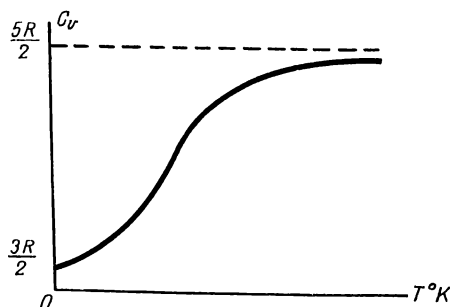


Fig. 84

If such an interval exists, the energy of a mole of gas and the thermal capacities of this quantity of gas are expressed by the following simple formulas:

$$U = 3RT,$$

$$c_v = 3R,$$

$$c_p = 4R.$$

The doubling of the internal energy and  $c_v$  with respect to a monatomic gas may be explained in the following manner. A polyatomic molecule has six degrees

of freedom, while a monatomic molecule has three. Since there are twice as many degrees of freedom, there is twice as much internal energy. To be sure, there is nothing self-evident about this statement. However, we find support for this viewpoint when we consider a gas consisting of diatomic molecules. Since a diatomic molecule is a system consisting of two particles, it possesses five degrees of freedom (see p. 32). If the internal energy is indeed proportional to the number of degrees of freedom, then for a gas consisting of diatomic molecules the following formulas should be valid:

$$U = \frac{5}{2} RT, \quad c_v = \frac{5}{2} R \quad \text{and} \quad c_p = \frac{7}{2} R.$$

Experiments show that in the temperature range in which the thermal capacity remains unchanged these formulas are quite applicable. The internal energy of one mole of a diatomic gas at a room temperature of 300 K is 1,500 cal = 6,250 J.

A typical dependence curve for thermal capacity over a broad interval of temperatures is illustrated in Fig. 84.

## Sec. 71. STATISTICAL DISTRIBUTION

Numerous events occur that cannot be predicted. These are called random events. The height of a young man appearing for military service, the number of pedestrians passing a particular crossing during certain hours, and the number of winning tickets in a loan lottery falling on each series of one hundred bond numbers are all examples of random events. The results obtained by observing numerous events of a single type, for example, measuring the height of a large number

of young people, counting the number of pedestrians per minute over a large number of days, or analysing the number of winning tickets for a large number of loan lotteries, may be summarised in the form of a so-called distribution curve. In the case of the height of a person, the data may be processed in the form of numbers indicating the number of men called up for military service whose heights are between 170 cm and 171 cm, between 171 cm and 172 cm, etc. Thus, the probability of observing a person among those called up who has precisely a given height (e. g., 171.34 cm) is practically equal to zero. Therefore, it is better to refer only to the number of men called up having a height lying in a particular interval.

In the case of the analysis of the prize list, the distribution curve may be constructed on the basis of the data for the number of series of one hundred bonds for which there was not a single winner, for which there was one winner, two winners, etc.

If we construct a graph, plotting the random quantity (e.g., height, number of pedestrians or number of winners) along the horizontal axis and the number of random events (e.g., number of people having a height lying in a particular interval, the quantity of cases of a given number of winners per one hundred numbers, etc.) along the vertical axis, the obtained curve is a distribution curve. An example of such a curve is shown in Fig. 85.

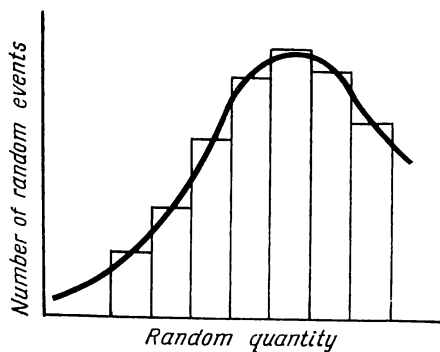


Fig. 85

The curve is drawn through the mid-points of the tops of the rectangles. Each rectangle has an area equal to the number of times a random event occurred for the quantity lying in the given interval.

The remarkable feature of distribution curves is their reproducibility. If we construct distribution curves analysing the height of young men called up for military service for a number of years, it will be seen that the curves are entirely similar. This similarity will not be found if we study the height distribution curves constructed on the basis of a small number of measurements. As we increase the number of measurements on which each curve is based, the curves for different years will become more and more similar. The same holds true for the distribution curves of all events, provided the events are random and the conditions of the obtained curves do not change.

We call the distribution law of one or another quantity a *statistical law*. It is more accurately given the greater the number of events used to determine each ordinate of the curve.

Naturally, knowing the distribution curve does not enable us to predict the number of a bond that will win in the next lottery. However, we can say, for example, what portion of the series consisting of one hundred numbers each will have one winner. The greater the number of bonds used in the analysis, the greater the accuracy of this prediction.

In view of the large number of molecules contained in very small volumes of matter, all kinds of statistical predictions about the behaviour of molecules are made with particularly high accuracy. A distribution curve of one or another random quantity plotted for the molecules of a substance will be reproduced with tremendous accuracy because each "rectangle" of the distribution curve corresponds to thousands of millions of molecules.

## Sec. 72. BOLTZMANN'S LAW

Certain ideas about the distribution of molecules follow immediately from the chaotic nature of thermal motion. This applies to the velocity distribution of the molecules according to direction and to the volume distribution of the molecules for the case when no forces act on the gas. However, there are numerous cases when the consequences of the assumption regarding the chaotic nature of thermal motion are not evident in advance.

First, there is the question of the distribution of the molecules according to speed. What percentage is moving rapidly, and what percentages are moving with average or slow speeds? Then, there is the problem of determining how a uniform density distribution of molecules changes when the gas is placed in a field of force, say in a gravitational field—or, if the molecules have electric or magnetic properties, in an electric or magnetic field. Boltzmann's law, which may be derived by means of the theory of probability, gives the answers to these and similar questions.

Let us consider a small volume of space—a cube at point  $x, y, z$  whose sides are  $\Delta x, \Delta y, \Delta z$ . Assume a considerable number of molecules to be contained in this cube. We shall consider those molecules having velocity components in the ranges from  $v_x$  to  $v_x + \Delta v_x$ ,  $v_y$  to  $v_y + \Delta v_y$  and  $v_z$  to  $v_z + \Delta v_z$ . The magnitudes of  $\Delta v_x, \Delta v_y$  and  $\Delta v_z$  are such that a large number of molecules are contained in the indicated interval of velocities. This is necessary in order to be able to apply the laws of statistical physics to these small volumes (physically, infinitesimal volumes). In the future, we shall say that such molecules have coordinates *in the neighbourhood of*  $x, y, z$  and velocities *in the neighbourhood of*  $v_x, v_y, v_z$ . We repeat, to speak of a quantity of molecules that have *exactly* a given velocity is impermissible, for the probability of encountering such a molecule is infinitely small. Since the kinetic energy of a molecule is determined by the value of the velocity and the potential energy of a molecule in an external field depends on the coordinates of the molecule in space, all the molecules segregated by us have, practically, one and the same energy  $\mathcal{E}$ .

Boltzmann's law, based on considerations developed in courses on theoretical physics, gives a general expression for the number of molecules whose coordinates are in the neighbourhood of  $x, y, z$  and velocities in the neighbourhood of  $v_x, v_y, v_z$ . This number is

$$\Delta n = A e^{-\mathcal{E}/kT} \Delta x \Delta y \Delta z \Delta v_x \Delta v_y \Delta v_z,$$

where  $A$  is a constant that may be determined for a concrete problem,  $T$  is the absolute temperature and  $k$  is Boltzmann's constant.

The energy in the exponent is equal to the sum of the kinetic energy of translation of the molecule and the potential energy of the molecule in the external field:

$$\mathcal{E} = \frac{mv^2}{2} + U. \text{ Hence,}$$

$$\Delta n = A e^{-\frac{\frac{mv^2}{2} + U}{kT}} \Delta x \Delta y \Delta z \Delta v_x \Delta v_y \Delta v_z.$$

This formula also applies to the case when the molecule possesses other forms of energy as well, for example, rotational and vibrational. These components of the energy must then be included in  $\mathcal{E}$ .

Boltzmann's law or, as it is also called, *Boltzmann's distribution*, shows that the largest energy corresponds to the lowest number of particles whose velocities and coordinates lie in the given interval.

We shall apply Boltzmann's law to the solution of important problems related to the height distribution of particles and the velocity distribution of molecules.

### Sec. 73. DISTRIBUTION OF PARTICLES WITH RESPECT TO HEIGHT IN A GRAVITATIONAL FIELD

If, in a liquid, there are a large number of small particles that are heavier than the liquid and do not dissolve in it, at first glance it may appear that sooner or later these particles must fall to the bottom. This, however, does not occur—but it would if there were no thermal motion.

Thus, the force of gravity attracts the particles downwards, but the chaotic thermal motion, an inherent property of all particles, will continuously impede the action of the gravitational force. A particle moving downwards may experience a collision on the way that hurls it back upwards. It again begins to move downwards and again a collision may hurl the particle upwards or sidewise. While some particle may succeed in reaching the bottom of the vessel, another particle, on the other hand, may be raised from the bottom by random impacts and brought to the upper layers of the liquid by random impulses. It is quite understandable that as a result some nonuniform distribution of particles is established. In the upper layers there will be the least number of particles, while at the bottom of the vessel there will be the greatest number. The heavier the particles and the lower the temperature, the more will the height distribution of particles be "compressed toward the bottom".

The quantitative aspect of this interesting phenomenon, occurring for all particles located in a gravitational field (molecules of a gas or particles of an emulsion suspended in a gas or liquid) becomes clear from Boltzmann's law. We may rewrite the exponential factor in the formula for the Boltzmann distribution in the form

$$e^{-\frac{mv^2}{2kT}} e^{-\frac{mgh}{kT}};$$

$U$ , the potential energy of gravitation, has been replaced by the expression  $mgh$ . Now, we are interested in the number of molecules (of all velocities) located at a height between  $h$  and  $h + \Delta h$ . It is

$$\Delta n = n_0 e^{-\frac{mgh}{kT}} \Delta h.$$

Here, the coefficient of proportionality  $n_0$  is simply the specific number of particles  $\frac{\Delta n}{\Delta h}$  at  $h = 0$ . Fig. 86 shows how the number of particles decreases with increasing height.

The form of this relation confirms the correctness of the assertion made above that the greater the mass of the particles and the lower the temperature, the more rapidly does the curve fall. It is also evident from the curve that its rate of de-



Fig. 86

crease depends on the gravitational acceleration. On different planets, the distribution of particles with respect to height will differ.

According to the above formula, at least a small number of molecules exists at every height above the Earth's surface. This means that molecules may recede from the Earth and fly into space, for it is not excluded that as a result of random collisions one or another molecule will attain a velocity of 11.5 km/sec, which is sufficient, as we know, to escape from the Earth's gravitational pull. It may, therefore, be stated that the Earth is gradually losing its atmosphere. However, calculation of the rate of dispersion of the atmosphere shows that it is negligible. During the entire existence of the Earth, an insignificant amount of air has been lost. The situation is different as regards the Moon, where the velocity required to overcome gravity is  $\sim 2$  km/sec. Such a small velocity is very easily attained by molecules and as a result the Moon has no atmosphere.

The formula giving the number of particles as the function of height may be rewritten for the density of a gas or for the pressure of a gas. Since the gas pressure is proportional to the number of particles in a unit volume, the formula may be written in the form

$$p = p_0 e^{-\frac{mgh}{kT}}.$$

Here,  $p_0$  is the pressure at zero level. This formula is called *the barometric formula*. It is used by meteorologists measuring atmospheric pressure at high altitudes to reduce the results of their measurements to "sea level".

It is necessary to note yet another important application of the formula for the distribution of particles with respect to height, which was used for the experimental determination of Avogadro's number by Perrin, the French scientist. In accordance with the conditions of the experiment, Perrin had to somewhat modify the formula for the distribution of molecules with respect to height. He studied an emulsion obtained by dissolving gutta gamba (a variety of resin) in water. Using a microscope, an entire mound of spherical granules could be observed in the emulsion. Perrin used a centrifuge to sort the gutta gamba granules according to size. Several months of labour yielded 20-30 g of gutta gamba granules having a diameter of 0.74 microns. The density of gutta gamba is  $D = 1.495$  g/cm<sup>3</sup>, i.e., the mass of one grain was equal to  $7 \times 10^{-14}$  g. Exact determination of the dimensions of the grains was no easy task. Perrin made this determination using three independent methods:

(1) The length of a chain of several dozen contiguous grains was determined under a microscope.

(2) The weight of several thousand grains was measured and the dimension calculated from the known density of gutta gamba.

(3) Stokes' formula (see p. 166) was used to determine the dimension from observations on the velocity with which a cloud of grains sinks in an emulsion. It was assumed that, in accordance with Archimedes' principle, a grain sinks under the action of the force  $\frac{4}{3}\pi r^3(D-d)g$ , where  $d$  is the density of the liquid and  $r$  is the radius of the grain. When the grain sinks uniformly, this force is balanced by the force of viscous friction calculated by Stokes' formula. From this condition, it is easy to determine  $r$ .

There was close agreement between the results of all three methods. This signified that the effective weight of a microscopic granule floating in a liquid may be written in the form  $mg\left(1 - \frac{d}{D}\right)$ . Recalling that  $k = \frac{R}{N}$ , we obtain the following barometric formula for an "atmosphere" of gutta gamba grains floating in water:

$$n = n_0 e^{-\frac{Nmg h}{RT}\left(1 - \frac{d}{D}\right)}.$$

The experiment reduced to the determination of the ratio of concentrations  $n$  at equal levels. This was accomplished by focussing the microscope on sufficiently thin layers of the emulsion and calculating the number of particles in the field of vision for equal intervals of time. Perrin changed the viscosity of the emulsion by a factor of one hundred and observed that the ratio of concentrations exactly agreed with the barometric formula. Substituting the values of  $n_0$ ,  $n$ ,  $h$ ,



$m$ ,  $d$ ,  $D$  and  $T$ , it was possible to determine  $N$ . It turned out that, in spite of the large changes in the viscosity of the emulsion and the dimensions of the grains,  $N$  determined in this manner agreed excellently with the values predicted by the kinetic molecular theory. Perrin obtained  $6 \times 10^{23} \leq N \leq 7 \times 10^{23}$ , while according to modern data  $N = 6.0225 \times 10^{23}$ . This was highly reliable evidence that the Boltzmann distribution according to energy is applicable even to particles having a gram molecule (mass of  $N$  particles) equal to 50,000 tons!

#### Sec. 74. VELOCITY DISTRIBUTION OF MOLECULES

The velocity distribution of molecules, first determined theoretically by Maxwell, an outstanding English physicist, may be considered to be a consequence of Boltzmann's law.

According to Boltzmann's law, the number of molecules whose velocities are in the interval from  $v_x$  to  $v_x + \Delta v_x$ ,  $v_y$  to  $v_y + \Delta v_y$ , and  $v_z$  to  $v_z + \Delta v_z$  is

$$\Delta n = C e^{-\frac{mv^2}{2hT}} \Delta v_x \Delta v_y \Delta v_z.$$

It is implied that we are interested in the velocity distribution in a small volume of gas and that the space distribution of the molecules is taken into account by the constant factor  $C$ , which does not interest us at the moment.

The above formula takes account of the distribution of the molecules with respect to the magnitudes as well as the directions of the velocities. However, we already know the distribution with respect to the directions—the number of molecules moving in one or another direction must be the same for complete randomness in the molecular motion. We are interested in the number of molecules having a speed from  $v$  to  $v + \Delta v$ , where

$$v = \sqrt{v_x^2 + v_y^2 + v_z^2}.$$

If we construct a three-dimensional diagram, along whose axes  $v_x$ ,  $v_y$ ,  $v_z$ , the projections of the velocities of the molecules, are plotted, and consider this space to be divided into infinitely small cubes of volume  $\Delta v_x \Delta v_y \Delta v_z$ , the data on the velocity distribution of molecules may be simply represented as the numbers of molecules contained in a cube. Boltzmann's formula gives us the number of molecules for each one of the cubes. However, examining the formula, we see that the number of molecules is the same for all cubes located within a spherical shell of radius  $v$  to  $v + \Delta v$ , for only the absolute value of the velocity enters in the exponential factor of the formula. The number of molecules having velocities in the range from  $v$  to  $v + \Delta v$  is proportional to the volume of the spherical shell, i.e.,  $4\pi v^2 \Delta v$ . Thus, if the number of molecules contained in one cube is equal to

$$C e^{-\frac{mv^2}{2hT}} \Delta v_x \Delta v_y \Delta v_z,$$

the number of molecules contained in the spherical shell, i.e., possessing velocities in the range from  $v$  to  $v + \Delta v$ , is represented by the formula

$$\Delta n = C e^{-\frac{mv^2}{2hT}} 4\pi v^2 \Delta v.$$

What then is the nature of this dependence? At  $v = 0$  and  $v = \infty$ , the number of molecules is equal to zero. It is evident that the curve must have a maximum. Let us determine in the usual manner the maximum of the factor preceding  $\Delta v$ . Taking the derivative of this expression and setting it equal to zero, we obtain

$$\frac{d}{dv} (e^{-\frac{mv^2}{2hT}} v^2) = 0.$$

Hence, the value of the velocity for which the distribution function has a maximum is

$$c = \sqrt{\frac{2kT}{m}}.$$

What can be said about this velocity? Since the number of molecules having the velocity  $v$  are plotted along the ordinate of the distribution curve,  $c$  is a peculiar boundary. Molecules moving with velocities greater or less than  $c$  are encountered less often than molecules of velocity  $c$ .

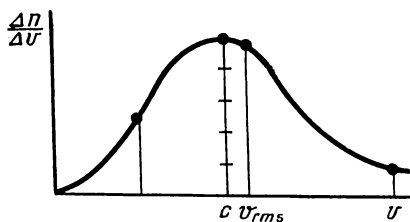


Fig. 87

This velocity is called the most probable. The velocity distribution curve for gas molecules (Maxwell distribution) is shown in Fig. 87.

It is interesting to compare the formulas for the most probable velocity and the r-m-s velocity:

$$c = \sqrt{\frac{2kT}{m}}, \quad \text{and} \quad v_{rms} = \sqrt{\frac{3kT}{m}}.$$

We see that the r-m-s velocity is greater than the most probable. The reason for this is evident from the form of the distribution curve. Since the curve extends far to the right, the root-mean-square velocity is displaced in that direction.

Let us cite several figures characterizing the velocity distribution of gas molecules. The number of molecules with velocities close to the most probable velocity  $c$  is 1.1 times larger than the number of molecules with velocities close to the root-mean-square value, 1.9 times larger than the number of molecules with velocities close to  $0.5c$ , and 5 times larger than the number of molecules with velocities close to  $2c$  (see Fig. 87).

## Sec. 75. MEASUREMENT OF THE VELOCITIES OF GAS MOLECULES

Even though the law of molecular velocity distribution is based on exceptionally well-founded theoretical grounds, whose validity is confirmed by a large number of physical facts, it is interesting to subject the distribution formula to direct experimental verification.

The velocity of gas molecules can be measured in a volume only by indirect means. If a molecule radiates light, the velocity of its motion affects the width of the spectral lines (Doppler effect).

Direct means of measurement require molecular beams. For this purpose, a long tube of large diameter is partitioned by two shutters having very small apertures. The gas is placed in an end compartment, whereupon the molecules begin, at first, to penetrate into the middle compartment and will sometimes even reach the compartment at the other end. Clearly, only those molecules whose vector velocities are directed along the axis of the tube when passing through the first aperture can traverse the entire length of the tube. Thus, a molecular beam is separated out from the gas. The velocities of the beam molecules all have the same direction. It is evident, however, that due to the random motion of the molecules, the distribution with respect to speed will be the same as for the molecules of any other direction of motion.

To measure the velocities of the beam molecules, we can resort to an arrangement reminiscent of an apparatus used for measuring the velocity of a bullet.

Such an apparatus has two cardboard disks rigidly fixed on a shaft and rotate about it with a velocity  $\omega$ . If the bullet travels parallel to the axis of rotation, the disks will be consecutively pierced at two points displaced in azimuth by an angle  $\varphi$  with respect to each other. This angle corresponds to the rotation of the system while the bullet traversed the distance  $l$  between the disks. The rotation time for angle  $\varphi$  is equal to  $\frac{\varphi}{\omega}$ . Hence, the velocity of the bullet is

$$v = l : \frac{\varphi}{\omega} = \frac{l\omega}{\varphi}.$$

Since molecules cannot pierce disks, the analogous experiment for molecular beams is performed with disks in which slits are cut along radii. The angular distance between the slits is equal to  $\varphi$ . Clearly, molecules of velocity  $v$  can pass through two slitted rotating disks only for a specific angular velocity  $\omega$ , satisfying the condition  $\frac{\varphi}{\omega} = \frac{l}{v}$ . Thus, by varying  $\omega$ , we can filter the molecules according to their velocities, collect molecules having the same velocities and measure their relative quantities.

The velocity distribution formulas discussed above, and hence the formulas for the r-m-s and most probable values of molecular velocities, have been verified by numerous experiments.

## Sec. 76. PROBABILITY OF A STATE

Let us consider a box divided into two equal parts by a partition in which an aperture has been cut. If there are molecules of gas in the box, they may be transferred from one half of the box to the other as the result of random collisions with the walls of the container and with each other.

In spite of the fact that the molecular motion is completely haphazard, a method exists for predicting how many molecules will be in the left half of the box and how many in the right half. This method is based on the application of the theory of probability.

If there were one molecule in the box, the chances or, as we say, the probability, that the molecule is in the right-hand portion is the same as that it is in the left-hand portion. Since, in all, there are two possible cases (the molecule is either in the left-hand or in the right-hand portion), and we are interested in the realisation of one of these cases, the probability of the molecule being in one half of the box is said to be equal to  $\frac{1}{2}$ . Now, assume that there are two molecules in the box, designated by the figures 1 and 2. In all, there are now four possible dispositions—both molecules on the left, both on the right, molecule No. 1 on the left and No. 2 on the right and, finally, No. 2 on the left and No. 1 on the right. We are interested in the probability of finding two molecules on the left. This is one case out of four possible ones, so that the probability is equal to  $\frac{1}{4}$ , i.e.,  $\left(\frac{1}{2}\right)^2$ . For three molecules, the situation is as follows:

left	1, 2, 3	0	1, 2	1, 3	2, 3	3	2	1
right	0	1, 2, 3	3	2	1	1, 2	1, 3	2, 3

Clearly, the probability of all three molecules being on the left is  $\frac{1}{8}$ , i.e.,  $\left(\frac{1}{2}\right)^3$ . It is not difficult to see that, for the case of  $N$  molecules, the probability of all

the molecules being in one part of the box is equal to  $\left(\frac{1}{2}\right)^N$ . Whenever another molecule is added, it is always possible to place it either in the left-hand or in the right-hand part. Therefore, with each newly added molecule, the probability of the molecules being in one half of the container is obtained by dividing the preceding probability by two.

When the number of molecules is still no more than one hundred, the quantity  $\left(\frac{1}{2}\right)^N$  is already so small that we need no longer take account of the possibility that all the molecules will be located in one half of the container. However, the number of molecules in a cubic centimetre of gas is not one hundred, but about  $10^{20}$ . If the container is considered to be divided into two parts, the probability that the molecules will all turn up in one half of the vessel is equal to  $\left(\frac{1}{2}\right)^{10^{20}}$ . By

taking the logarithm, this number may be converted into the form  $10^{-3 \times 10^{19}}$ . To put this number in decimal form,  $3 \times 10^{19}$  zeros must be written! A person writing at a rapid speed of three zeros per second will require  $10^{19}$  sec to write this number. This is equivalent to 300,000 million years, which is ten times the amount of time our solar system has been in existence.

Let us return to the table for the disposition of three molecules. Only for one disposition out of eight do all the molecules turn up on the left. Every other disposition is also encountered one time out of eight. It should be remembered, however, that the molecules are arbitrarily numbered and there is no way of differentiating a disposition in which Nos. 1 and 2 are on the left from one in which 2 and 3, or 1 and 3, are on the left. Thus, compared to the one disposition in which there are three molecules on the left, there are three dispositions in which there are two molecules and the same number of dispositions in which there is one molecule on the left. Therefore, the probability of some characteristic distribution existing, regardless which particular molecules produce it, may be measured by the number of dispositions that can produce the distribution. The greater this number, the more frequently will such a distribution be encountered, i.e., the more probable will this distribution be.

This example brings us to the concept of probability of a state of a body.

At each instant the atoms of which the body is made up have certain coordinates and velocities. This instantaneous structure will be called a microstate.

Any body which is in a state of equilibrium with the surrounding medium retains all its properties, but, nevertheless, it does not remain in one and the same microstate. Due to thermal motion of the particles, the body continuously changes its microstates. In gases these changes are caused by translational (or progressive) motion, vibration, and revolution of the particles; in liquids, microstates are changed owing to vibrations of the particles and their transition from one surrounding to another; in solids—mainly due to vibrations. In any case, the body is in a dynamic equilibrium.

Passing from one microstate to another, the body will repeatedly return to one and the same state. Some of these states occur more frequently, others more rarely, as it is clear from the above considered example.

If during a long period of time  $T$  a body lived in a certain microstate for a time  $\Delta t$ , then  $\Delta t/T$  is the probability of this microstate.

The probability of a microstate is expressed by a simple formula derived by J. W. Gibbs:

$$w = Ae^{-\mathcal{E}/kT}.$$

where  $\mathcal{E}$  is energy. The constant  $A$  takes into account the number of permutations through which a microstate can be realised. With equal  $A$ 's the probability of a microstate is determined by its energy.

The form of the above formula coincides with that of the Boltzmann law. What is then the relationship between these two laws? The Boltzmann formula considers a large number of molecules (or bodies) at one instant and tells us about the energy distribution of these molecules (bodies), whereas the Gibbs formula is applied to one body (or molecule) which is "watched" by us for a long time, and yields information concerning energy distribution of this body in time. Of course, this is not a random coincidence.

As it was already stated, the discreteness of a body state (quantum state) belongs to the principal laws of nature. Therefore, we may speak of the number of microstates by which a given microstate of the body is realised. This number is called the statistical weight of a macroscopic state (another term for thermodynamic probability).

The thermodynamic probability  $W$  is uniquely associated with thermodynamic functions of a body. It is easily understood that the statistical weight of a state increases with temperature; it increases during the processes of melting, evaporation, etc. We may assert that the greater the freedom of motion of the particles the body is made up of, the higher the thermodynamic probability of the state.

We can obviously imagine the relationship between the observable (microscopic) quantities and the probability of microstates. It is clear that the values obtained for a microscopic quantity are mean values of those attained by this quantity for microstates. If, for instance, for the  $n$ th microstate the energy is equal to  $E_n$ , then the mean (observed) energy

$$E = w_1 E_1 + w_2 E_2 + w_3 E_3 + \dots$$

Of course, the probabilities  $w_n$  must be normalised to unity ( $\sum w_n = 1$ ).

## Sec. 77. IRREVERSIBLE PROCESSES FROM THE MOLECULAR VIEWPOINT

The example of the previous article shows quite clearly that the state in which the molecules are distributed "uniformly" is the most probable. Any deviation from "uniformity"—displacement of one portion of the molecules to the left side of the container, disposition of the faster molecules on the left, directed motion of a large portion of the molecules, in short, any deviation from haphazardness in the space and velocity distribution of the molecules—results in a decrease in the probability of the state. This conclusion enables us to comprehend the kinetic-molecular nature of the irreversibility of actual processes.

As was established above, the second law of thermodynamics for irreversible processes, i.e., the law of increasing entropy in thermally isolated systems, is a generalisation of experimental experience on the impossibility of certain processes. Thus, heat cannot be transferred from a cold body to a hot one without compensation, a body cannot acquire kinetic energy merely at the expense of a decrease in the internal energy of the surrounding medium, and a gas may expand of itself but not contract.

The existence of irreversible processes is a peculiarity of molecular phenomena. For a purely mechanical phenomenon, i.e., a process without friction, the process may always be reversed. When a pendulum moves to the right, it passes in reverse order through all the states passed in moving to the left. A billiard ball rebounding from the side of the table in some direction or other will, in turn, rebound from an elastic wall placed in its path and retrace in reverse order the entire path tra-

versed in the "forward" direction. The complete equivalence of "forward" and "backward" is evident for purely mechanical processes. Why then do molecular processes, which we have considered as the totality of the movements of the molecules, not possess the property of reversibility? There is only one reason. In all irreversible processes, the probability of the state increases. A reversible process is a conceivable process, i.e., its implementation is in principle possible, but for the time available to a person for observation such a process is, practically speaking, improbable.

This is not difficult to show for any irreversible process. Heat transfers from a hot body to a cold one, but not vice versa. In the case of gaseous bodies, such a process may be visualised as a mixing of fast molecules with slow ones. The reverse process cannot occur according to the random law, for it would constitute a sorting out of fast and slow molecules, i.e., a transition to a more orderly state.

For the same reason, using a shovel, we can quite rapidly mix the contents of two sacks of different grain. However, we can continue to mix the contents of these two sacks endlessly without the grain separating in such a manner that one of the grain varieties appears above and the other below. It should be realised, moreover, that the number of kernels in the sacks is immeasurably less than the number of molecules in a cubic millimetre.

It is also easily seen that the reverse process of spontaneous expansion of a gas is completely improbable. If in the partitioned box considered above there is gas on the left and vacuum on the right, both parts of the box will be uniformly filled with gas within a short time. In principle, it could occur that all the molecules turn up together again on the left. However, the probability of such an event is extremely small. Its value has been shown to be equal to  $\left(\frac{1}{2}\right)^N$ .

No matter which irreversible process we consider, the result will always be the same—each irreversible process is accompanied by an increase in the probability of the state.

Thus, there are two quantities that increase during irreversible processes—the entropy  $S$ , with which we are already familiar, and the probability of the state  $W$ , which we have just discussed. It is natural that these two physical quantities should be related. Boltzmann showed that such a relation does, in fact, exist. The formula given by him has the form  $S = k \ln W$ , i.e., the entropy is proportional to the logarithm of the probability of the state.

The second law of thermodynamics thus acquires still another formulation: in reversible processes, the probability of the state does not change, while in irreversible processes (we are referring to closed systems) the probability of the state increases.

## Sec. 78. FLUCTUATIONS. LIMITS TO THE APPLICATION OF THE SECOND LAW

All physical properties remain unchanged if the space and velocity distributions of the molecules do not change. In principle, the distribution of the molecules of a substance may change with time. However, as we have indicated above, the most probable distributions stand out so sharply that deviations from these distributions must be considered to be very rare events. The physical characteristics corresponding to this most probable distribution are called the *average characteristics*. Practically, the deviation of a measured physical characteristic from its average value for systems having large numbers of molecules is impossible to observe. This is the situation when the physical properties are being considered for volumes containing large numbers of molecules. However, if the number of

particles in a system is small, it is also possible to observe rarer space and velocity distributions of molecules. The values of the physical characteristics corresponding to these rarer distributions differ from the average values. These deviations of the physical characteristics from their average values, which occur in systems having a relatively small number of particles, are called *fluctuations*. All properties of volumes containing small numbers of molecules—e.g., temperature and pressure, thermal capacity and thermal conductivity—are subject to fluctuations about the average values.

This question may be approached somewhat differently. If a tiny mirror suspended from a thin string is placed in a gaseous medium, then, from the macroscopic standpoint, the pressure of the gas acting on the mirror cannot manifest itself, for the forces act on all sides equally. In principle, from the molecular standpoint, the changes in momentum due to the impacts of the molecules on the mirror do not necessarily have to balance for the different portions of its surface. A light mirror may thus begin to execute fluctuational vibrations. As was stated above, for any particle (molecule, Brownian particle, pea), the energy of thermal, random motion is equal to  $\frac{1}{2} kT$  for one degree of freedom of motion. And this is the energy, on the average, falling on the mirror. On the other hand, the work of rotating the string by an angle  $\Delta\varphi$  is equal to  $M\Delta\varphi$ . Therefore, angular deflections of the order of magnitude  $\Delta\varphi \approx \frac{kT}{M}$  will occur quite often.

Such fluctuations are indeed observed and their measurement may be used for the experimental determination of Boltzmann's constant and, hence, of Avogadro's number.

Fluctuational effects limit the accuracy of measurements. A pointer, mirror, or any other indicating device is subject to fluctuations. At room temperature, the accuracy limit in energy units is about  $10^{-20}$  J. In the construction of many instruments such accuracy has not yet been achieved, but in some of the best measuring devices it already has.

Fluctuations limit the applicability of the second law of thermodynamics. They represent processes in which the system passes from a more probable state to a less probable one, i.e., processes in which the entropy decreases.

This is excellently illustrated by Brownian movement, where the pressure fluctuations occur in small volumes and affect individual particles. Due to these random pressure vibrations, a particle may, for example, be impelled upwards. However, motion against the force of gravity requires work. In this case, the work is performed at the expense of the random, thermal motion of the molecules, i.e., only at the expense of the internal energy of the substance, which is at complete variance with the second law of thermodynamics.

Although phenomena for which the entropy decreases occur at times in individual small volumes, i.e., the second law of thermodynamics is contradicted, the system as a whole will always obey this law. Due to the randomness of the events, the number of processes occurring at the expense of the internal energy will be exactly equal to the number of processes occurring in the reverse direction. It can be rigorously proved that any attempt to create a perpetual engine of the second kind by "selecting" in individual small volumes processes that contradict the second law will end in failure.

The second law of thermodynamics has a limit at the other end of the scale as well. In addition to being inapplicable to systems having a very small number of particles, it loses validity for systems having an infinitely large number of particles, namely, the universe or any of its infinitely large components. As was ex-

plained above, the essence of the second law of thermodynamics consists in the fact that the number of equilibrium states is overwhelmingly greater than the number of nonequilibrium distributions. However, for the universe, which consists of an infinitely large number of particles, this statement loses its meaning, for the number of equilibrium states and the number of nonequilibrium states are both infinitely large. Consequently, for the universe as a whole, one cannot speak of the differences in the probability of states.



# Processes of Transition to Equilibrium

## Sec. 79. DIFFUSION

A body interacting with its medium changes its state so as to come into equilibrium with the bodies surrounding it: its internal energy tends to a minimum, while its entropy increases and becomes a maximum when equilibrium is established. These two tendencies are usually conflicting. As a result it is difficult to predict the effect when both energy and entropy are capable of changing. Let us now consider the phenomena of diffusion, thermal conductivity and internal friction occurring in closed systems. In other words, we are concerned with equalisation of the concentrations, temperatures and velocities of some parts of a body with respect to others. (Naturally, equalisation of velocities is meaningful only for liquids and gases.) Since the energy cannot change in such systems, the transition to a state of equilibrium consists only in an increase in entropy.

The basic laws for the phenomena of diffusion, thermal conductivity and internal friction are very similar. Let us first consider diffusion processes. It is immaterial whether we deal with the equalisation of the concentration of a gas or a liquid. Our discussion will even be valid for solid solutions (see p. 490), since in this case too the tendency to maximum entropy makes the atoms or molecules of a substance intermix so that a single concentration is established in all parts of the body.

Let us consider two physically close, infinitely small volumes of a substance whose concentrations of diffusing atoms (or molecules) are  $c$  and  $c + dc$ . If these two volumes are a distance  $dx$  apart, the ratio  $\frac{dc}{dx}$  gives the rate of change of concentration. This ratio is called *the concentration gradient*. If the  $x$ -axis is chosen so that its positive direction coincides with the diffusion direction, then  $\frac{dc}{dx}$  will be a negative quantity. Substance will migrate in the direction of lower concentrations.

This does not mean that all the molecules move in one direction in a continuous, uninterrupted stream. On the contrary, diffusion maintains to a considerable extent the haphazard features peculiar to molecular motion. The molecules move haphazardly in all directions, including the direction of greater concentration, but the probability of molecules being displaced in the "correct" direction is greater than for other directions. This means that through an area perpendicular to the flow more particles pass in the direction from greater concentration to less concentration than the other way round.

The basic law of diffusion states that the flow of matter  $\mu$ , i.e., the mass of matter passing through a unit area per unit time, is directly proportional to the negative gradient of the concentration:

$$\mu = -D \frac{dc}{dx}.$$

$D$ , the constant of proportionality, is called *the coefficient of diffusion*. The above relation is convenient since the coefficient of diffusion is, within broad limits, a constant for a given substance and medium.

In measuring the concentration and the flow of matter, the units should correspond. Thus, if the concentration is measured in grams per  $\text{cm}^3$ , the flow should be measured in grams per  $\text{cm}^2$  per sec. We see then that the dimensions of the coefficient of diffusion are completely determined and in the CGS system are expressed in  $\text{cm}^2/\text{sec}$ .

A decrease in concentration usually follows a sagging curve as shown in Fig. 88. If we are interested in the portion for which the decrease in concentration may be represented by a straight line, then

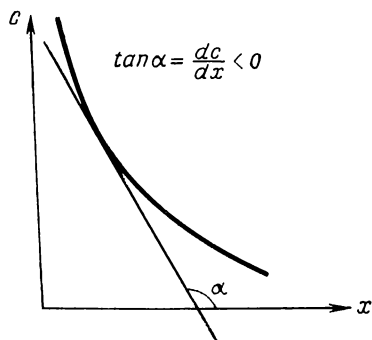


Fig. 88

$$\mu = -D \frac{c_2 - c_1}{x_2 - x_1},$$

where  $c_1$  and  $c_2$  are the values of the concentrations at points  $x_1$  and  $x_2$ , respectively.

The coefficients of diffusion vary within broad limits. For example:

(1) for gases at temperatures from  $0^\circ$  to  $15^\circ\text{C}$ :

hydrogen $\rightarrow$ oxygen,	$D = 0.778 \text{ cm}^2/\text{sec}$ ;
air $\rightarrow$ oxygen,	$D = 0.178 \text{ cm}^2/\text{sec}$ ;
air $\rightarrow$ carbon disulphide,	$D = 0.099 \text{ cm}^2/\text{sec}$ ;

(2) for solutions of blue vitriol diffusing in distilled water (where  $c$  is in gram equivalents per litre):

$c$	$D \text{ (cm}^2/\text{day)}$
0.1	0.39
0.5	0.29
0.95	0.23

## Sec. 80. THERMAL CONDUCTIVITY AND VISCOSITY

The temperature equalisation process is very similar to that of diffusion. If the temperature of a body differs at different points, the entropy is not a maximum. In order for equilibrium to be established, the average velocities of the molecules, and hence the temperatures, must become equalised.

If the temperatures at two neighbouring points separated by a distance  $dx$  are  $T$  and  $T + dT$ , the ratio  $\frac{dT}{dx}$  expresses the rate of temperature drop and is called *the temperature gradient*.

During the process of temperature equalisation, portions of the body having more energy give up energy to portions of the body having less energy. In a certain sense, heat "flows" from one portion to another. The amount of heat passing from one portion of the body to another through a unit area per unit time is called *the heat flow*  $q$ . Just as in the case of diffusion, one can assume that the heat flow is proportional to the negative temperature gradient. The greater the temperature difference, the more rapid the heat flow. The formula

$$q = -\kappa \frac{dT}{dx}$$

is convenient here too since the constant of proportionality  $\kappa$ , which is called *the coefficient of thermal conductivity*, is a constant for a given substance and does not depend on the magnitude of the heat flow. For a linear temperature drop, the

formula assumes the simplified form

$$q = -\kappa \frac{T_2 - T_1}{x_2 - x_1}.$$

It is not difficult to determine the dimensions of the coefficient of thermal conductivity. In the CGS system, this coefficient is measured in cal/(cm sec K). It is evident from the formula that  $\kappa$  is the heat flow per second through an area of 1 cm<sup>2</sup> when the temperature drops 1 K over a distance of 1 cm.

The values of the coefficient of thermal conductivity, just as in the case of the coefficient of diffusion, vary within broad limits.

For example:

(1) for solid bodies (0°–18°C): cork—0.00012, wood—0.0008, fused quartz—0.0033 and silver—1.06 cal/(cm sec K).

(2) for liquids: carbon disulphide (14°C)— $2 \times 10^{-4}$ , sulphuric acid 30% (32°C)— $62.4 \times 10^{-4}$ , mercury (0°C)—0.2 cal/(cm sec K).

(3) for gases (0°C): carbon dioxide— $3.4 \times 10^{-5}$ , air— $5.7 \times 10^{-5}$ , hydrogen— $40.6 \times 10^{-5}$  cal/(cm sec K).

The third phenomenon of this class relates to the equalisation of velocities. If a gas or liquid moves in some direction or other so that different layers of the substance move at different velocities, the motion is unstable. Sooner or later, the velocities must become equalised—the slower layers are accelerated and the faster ones are retarded. This phenomenon is also called *internal friction* or *viscosity*. It is typical for all substances except helium II (see Sec. 256).

Let us consider a liquid or gas moving along an  $x$ -axis. Assume that different layers of the fluid are moving at different velocities. Along the  $y$ -axis, perpendicular to the direction of flow of the liquid or gas, take two close points separated by the distance  $dy$ . The velocities of flow differ at these two points by  $dv$ . Thus, the ratio  $\frac{dv}{dy}$  is the gradient of the velocity and expresses the rate of change of the velocity as we move away from the surface of the fluid. To make this clear, consider a rapid stream. The velocity of flow is a maximum at the surface and gradually decreases as the bottom of the stream is approached.

If at some instant we eliminate the causes of the fluid motion, the velocities of the different layers will begin to be equalised in accordance with the law of increasing entropy. In order for such an equalisation to be possible, an internal frictional force must exist between the layers of the liquid or gas. The magnitude of this force per unit layer area is proportional to the gradient of the velocity, i.e.,

$$f = -\eta \frac{dv}{dy}.$$

Here,  $\eta$  is the *coefficient of viscosity* (or internal friction). Its dimensions in the CGS system are g/(cm sec). Such a unit is called a *poise* (P). In the SI system  $\text{N} \cdot \text{sec} \cdot \text{m}^{-2} = 10 \text{ P}$ .

The viscosity of different bodies varies within even broader limits than the two analogous coefficients considered above. For example:

(1) Solid bodies: glass (710°C)— $4.5 \times 10^{10}$ , glass (420°C)— $4 \times 10^{10}$ , lead (9°C)— $4.7 \times 10^{14}$ , ice (−14°C)— $8.5 \times 10^{12}$  P.

(2) Liquids: ethyl ether (25°C)—0.0022, water (20°C)—0.01, glycerine (0.8% water, 18°C)—13.93 P.

(3) Gases: hydrogen (0°C)— $8.49 \times 10^{-5}$ , air (0°C)— $17.19 \times 10^{-5}$  P.

It is interesting to note that hydrogen has one-half the viscosity of air and seven times its thermal conductivity. That is why hydrogen is used to cool powerful turbogenerators.

## Sec. 81. RATE OF EQUALISATION

We all know that the time it takes for equilibrium to be established varies within broad limits. The temperature of a red-hot piece of iron thrown into water and the temperature of the water become equalised rapidly. On the other hand, the temperatures of air and a hot brick become equalised slowly. Nitrogen diffuses almost instantaneously in oxygen, while the equalisation of the concentrations of a solution of blue vitriol takes many days. Similarly, the equalisation of velocities also varies within broad limits, depending on whether we are dealing with a gas or a viscous liquid.

A universal formula for the time of equalisation cannot be given, for the geometry of the object affects the time. A cooling body may have the form of a cylinder or a plate; a diffusing gas may initially be within a small spherical volume or distributed over some surface; and the internal friction may occur in pipes of various cross-section or in open reservoirs. The particular circumstances must be taken into account in each case and exact calculation of the value for the time of equalisation is a difficult mathematical problem. However, we can abstract from the geometric details and try to solve the problem in general form if we abandon the aim of obtaining an exact formula and are content to merely determine the proportionality between the physical quantities. In this connection, it is helpful to consider the dimensions of the physical quantities that should be related to each other.

Let us consider, for example, the phenomenon of diffusion. It is evident that  $t$ , the time of concentration equalisation, depends, in the first place, on the dimensions of the region in which the diffusion occurs (characteristic length  $L$ ) and the properties of the diffusing substance (characterised by the coefficient of diffusion  $D$ ). The diffusion equation has the form  $\mu = -D \frac{dc}{dx}$ . The dimensional equation is then

$$\frac{M}{L^2 T} = [D] \frac{M}{L^4}.$$

We see that  $T = \frac{L^2}{[D]}$ , i.e., the time of equalisation  $t = K \frac{L^2}{D}$  and does not depend on the concentration.

Therefore, the following conclusion may be drawn: every rigorous solution determining the time of concentration equalisation for diffusion processes will always yield the equation

$$t = K \frac{L^2}{D},$$

where  $K$  is a constant, dimensionless quantity depending on the geometric conditions of the problem. The quantity  $L$ , on whose square the speed of concentration equalisation depends, refers to the geometric dimension of the region in which the equalisation occurs. Thus, if the concentration becomes equalised within the limits of one centimetre in, say, 10 sec, then, within the limits of two centimetres, it will become equalised in 40 sec.

We can solve the problem of temperature equalisation in exactly the same manner. The basic relation of this phenomenon includes the following quantities: heat, coefficient of thermal conductivity, temperature and distance. However, the increment of heat per unit volume may be written in the form

$$dq = \rho c_p dT,$$

where  $c_p$  is the specific heat at constant pressure and  $\rho$  is the density (thus,  $c_p\rho$  is the thermal capacity of a unit volume). Therefore, the following quantities must be related to each other: temperature, length, time, density, thermal capacity and thermal conductivity. It is easily verified that the time  $\tau$  cannot depend on the temperature and can only be expressed by the other quantities as follows:

$$\frac{L^2\rho c_p}{\kappa}.$$

Thus, the time of temperature equalisation is expressed by the formula

$$t = K \frac{L^2}{\chi},$$

where  $\chi$  designates the combination of constants  $\frac{\kappa}{\rho c_p}$ . The quantity  $\chi$  is called *the thermometric conductivity* and has been introduced in order to put the formulas for the equalisation of concentration and temperature in similar form. The coefficients of diffusion and thermometric conductivity have the same dimensions and are completely analogous in the two equalisation phenomena considered.

We thus see how the cooling of a body is determined. The greater the density and thermal capacity, and the smaller the coefficient of thermal conductivity, the slower the process.

*Example.* Let us compare two rods of identical dimensions made of fused quartz and silver, respectively. For quartz,  $\kappa = 0.0033$  cal/(cm sec K),  $\rho = 2.65$  g/cm<sup>3</sup>, and  $c_p = 0.1844$  cal/(g K), i.e.,  $\chi = 0.676 \times 10^{-2}$  cm<sup>2</sup>/sec. For silver,  $\kappa = 1.06$  cal/(cm sec K),  $\rho = 10.5$  g/cm<sup>3</sup> and  $c_p = 0.0558$  cal/(g K), i.e.,  $\chi = 1.71$  cm<sup>2</sup>/sec. Thus, the equalisation of temperature in the quartz rod takes 253 times longer than in the silver rod.

Just as for diffusion, equalisation of the temperatures is characterised by a dependence on the square of distance, i.e., the time of equalisation is proportional to the square of the linear dimension of the region.

Without going through an analogous procedure, let us write the formula for the time of velocity equalisation for the various parts of a liquid or gas. It is not surprising that the form of this relation is similar, namely:

$$t = K \frac{L^2}{\nu}.$$

The coefficient  $\nu$ , which determines the rate of equalisation of the velocities, is equal to  $\frac{\eta}{\rho}$  and is called *the kinematic viscosity*.

*Example.* For water,  $\eta = 0.01$  P and  $\rho = 1$  g/cm<sup>3</sup>, i.e.,  $\nu = 0.01$  cm<sup>2</sup>/sec; for glycerine,  $\eta = 13.9$  P and  $\rho = 1.25$  g/cm<sup>3</sup>, i.e.,  $\nu = 11.1$  cm<sup>2</sup>/sec. This means that if a disturbance in glycerine is equalised in 0.1 sec, the same disturbance in water will be equalised in about 2 minutes.

## Sec. 82. STEADY PROCESSES

If a body is left undisturbed, the differences in temperature concentration and velocity of the various parts of the body will be equalised, to be sure, in accordance with the principle of increasing entropy. However, it is also possible to have a state of a body for which, over a prolonged period, the flow of heat or matter, or the velocity distribution of various parts of the body with respect to each other, remains unchanged. Processes of this type are called *steady* processes. Naturally, in the case of a steady process, the body is not in a state of equilibrium.

Under what conditions are such processes possible? Let us consider a metallic rod to which at each instant of time, a certain quantity of heat is supplied at one end of the rod, while the other end is in thermal contact with a cold body. The condition under which the temperatures along the rod will not change, i.e., the condition of constant temperature gradient along the entire path of heat flow, is that the quantity of heat absorbed by the cold body be exactly equal to the quantity of heat supplied by the hot body during the same period of time.

Under analogous conditions, a steady diffusion process is also possible. To create such a process, a certain quantity of matter must be supplied to one part of a body and the same quantity must be removed from another part. In this manner, a constant difference of concentration is maintained between two parts of the body.

A steady viscosity process may be obtained, for example, in the region between two coaxial cylinders rotating at different velocities. Since close to the solid surface the liquid or gas has the same velocity as the solid wall, a constant velocity gradient is created within the fluid.

Steady processes do not arise immediately. A certain amount of time must elapse for such processes to become established.

Let us assume that one end of a rod that transmits heat is placed in snow. At the initial instant, the temperature of the rod is equal to zero at all points. If the other end of the rod is then brought into thermal contact with boiling water, the temperature begins to rise throughout the rod, but, of course, not at the same rate at all points. Almost immediately, a high temperature is established at the end of the rod that is in contact with the boiling water. The temperature of the end of the rod that has been placed in snow will rise slowest. After a certain period of time, the temperature will cease rising at all points of the rod and a definite temperature distribution becomes established, i.e., the process becomes steady. The nature of the temperature distribution depends on the amount of heat supplied (or removed) per unit time.

In an electric iron heated by a spiral element, the highest temperature is in the central region and the temperature gradually drops towards the outer edges. Naturally, the air immediately surrounding the iron is hottest. With increasing distance from the iron, the temperature falls more rapidly owing to the low thermal conductivity of air.

In the case of small bodies in air or liquid, the temperature curve need not be considered in rough calculations, i.e., it suffices to deal with the temperature difference  $T - T_0$  between the body and the medium. The heat flow per unit time from the body to the medium may then be assumed to be proportional to this temperature difference:

$$q = k (T - T_0).$$

The coefficient  $k$  is called *the coefficient of thermal output* and is an important engineering quantity. In courses on heat engineering, values for this coefficient are determined and related calculations discussed.

Let us designate by  $P$  the power supplied to a body, e.g., electric power in the case of an electric iron. The condition for a steady process requires that

$$P = k (T - T_0).$$

Here,  $T$  is the body temperature established in this steady process:  $T = T_0 + \frac{P}{k}$ . It may vary considerably, depending on the power supplied and the conditions of heat exchange.

It is appropriate at this point to comment on the temperature indicated by a thermometer placed "in the sun". The thermometer is involved in the steady process of transferring solar heat to the surrounding air. Depending on the value of the coefficient of thermal output, the thermometer lying in the direct rays of the sun may indicate any value whatsoever. The temperature measured under such conditions is the temperature of the thermometer and is in no way an indicator of the weather.

We shall not consider the analogous diffusion problems.

### Sec. 83. MOTION IN A VISCOUS MEDIUM

Consideration of the dimensions of physical quantities helps to solve problems of tremendous practical importance. One such problem is the steady flow of a liquid or gas around an obstacle or, what amounts to the same thing, the steady motion of a body in a medium.

The most important problem concerns the resistance force experienced by a body in moving through a medium. This resistance force may depend on the body dimension  $L$ , the body velocity  $u$ , and properties of the liquid (or gas), namely, its density  $\rho$  and viscosity  $\eta$ . Other quantities should play no part in this process.

Let us first consider the dimensionless quantity comprised of  $L$ ,  $u$ ,  $\rho$  and  $\eta$ . It will be recalled that the kinematic viscosity,  $\nu = \frac{\eta}{\rho}$ , has the dimension  $L^2 T^{-1}$ . But the product  $Lu$  also has this dimension. Therefore,

$$Re = \frac{\rho Lu}{\eta} = \frac{Lu}{\nu}$$

is a dimensionless quantity. This quantity is designated as indicated and is called *the Reynolds number*. It can be shown that  $Re$  is, in effect, the only dimensionless combination of the indicated quantities. Other dimensionless quantities can only be functions of the Reynolds number, i.e.,  $f(Re)$ . If the motions of different bodies in different fluids lead to one and the same value for  $Re$ , the motions are said to be similar. A large technical field founded on the principle of similitude has developed. In this field, the characteristics of a phenomenon are determined on the basis of observations of a similar phenomenon occurring in a model.

Let us now return to the problem raised above, namely, finding the expression for the resistance force experienced by a body moving in a medium.

The dimensions of force are given by  $MLT^{-2}$ . This can be equated to the dimensions of the quantities with which we are dealing, since it cannot depend on any other quantities. Thus,

$$MLT^{-2} = [\rho]^\alpha [u]^\beta [L]^\gamma [\eta]^\delta$$

i.e.,

$$MLT^{-2} = M^\alpha L^{-3\alpha} L^\beta T^{-\beta} L^\gamma M^\delta L^{-\delta} T^{-\sigma}.$$

Hence,

$$\alpha + \delta = 1, \quad -3\alpha + \beta + \gamma - \delta = 1, \quad -\beta - \delta = -2.$$

Expressing  $\alpha$ ,  $\beta$  and  $\gamma$  in terms of  $\delta$ , we obtain

$$\alpha = 1 - \delta, \quad \beta = 2 - \delta, \quad \gamma = 2 - \delta.$$

Thus, in the most general case,  $F$  may be expressed in the form of a sum of terms, each of which has the indicated dimensions, i.e.,

$$F = A [\rho^{1-\delta} u^{2-\delta} L^{2-\delta} \eta^\delta] = \rho u^2 L^2 A \left[ \left( \frac{Lu\rho}{\eta} \right)^{-\delta} \right],$$

where  $A$  represents numerical coefficients. We have thereby demonstrated that the force must be given by the formula

$$F = K\rho u^2 L^2 f(\text{Re}).$$

This result has been obtained only by considering the dimensions! The function  $f(\text{Re})$  is not known and must be determined experimentally.

Definitive formulas for limiting cases may be obtained by simple reasoning. If the velocity is small,  $F$  must be proportional to the first power of the velocity  $u$ . For this to be true,  $f(\text{Re})$  must be equal to  $\frac{1}{\text{Re}}$  and, therefore,

$$F = K\eta uL.$$

The numerical value of the constant depends on the form of the body. For a sphere,

$$F = 6\pi\eta ur,$$

where  $r$  is the radius of the sphere. The last formula is known as Stokes' formula.

*Example.* A mercury globule ( $r = 0.53$  mm), sinking in glycerine with a velocity of 0.6 cm/sec, experiences a force of friction of about 8 dynes.

In the case of very large velocities, the fluid motion with respect to a body ceases to be steady. Eddies appear and the motion becomes turbulent. The body motion may be steady, but the fluid particles move more or less randomly. Owing to the intense nature of the disturbance, the transfer of motion from layer to layer ceases to depend on the viscosity. This can only occur if  $f(\text{Re})$  approaches a limit as the velocity increases. Therefore, for large velocities, the resistance force becomes proportional to the velocity squared:

$$F = K\rho u^2 L^2.$$

#### Sec. 84. COEFFICIENTS OF DIFFUSION, VISCOSITY AND THERMAL CONDUCTIVITY FOR GASES

The processes of equilibrium establishment in gases are closely related to the characteristics discussed in the previous article. Temperature, concentration and velocity equalisation of some parts of a gas with respect to others occur owing to the mixing of the molecules. The rate of this mixing is determined by the role played by collisions between particles. For example, in case of a large free path, the fast molecules quickly penetrate into the regions where the slow molecules are located and distribute themselves throughout the gas.

It is quite natural that the time of equalisation in all three processes should be of the same order of magnitude as the time between molecular collisions. This may be verified by theoretical calculations for particular cases, but we shall not concern ourselves with this problem.

Taking the equation for the equalisation time to be  $\tau = \frac{l}{v}$ , whereby a dimensionless constant of proportionality whose order of magnitude is usually equal to unity is omitted, we obtain on the basis of Sec. 81 perfectly identical expressions



for the coefficients of diffusion,\* kinematic viscosity and thermometric conductivity (assuming  $L \approx l$ ):  $D \sim \nu \sim \chi \sim \nu l$ .

The following table indicates the accuracy obtained:

air	hydrogen
$\nu = 0.13$	$\nu = 0.94$
$\chi = 0.18$	$\chi = 1.3$
$\nu l = 0.27$	$\nu l = 1.9$

These results should be considered good. Agreement within an order of magnitude cannot be viewed as accidental when it is recalled how greatly the quantities with which we are dealing vary.

Using the expression for the coefficient of thermal conductivity in terms of the thermometric conductivity, we obtain

$$\kappa \sim \rho \nu l c_p \sim \frac{m c_p \nu}{\sigma},$$

where  $m$  is the mass of a molecule.

In this formula,  $n$ , the number of molecules in a unit volume, has cancelled out. It follows, therefore, that the thermal conductivity of a gas does not depend on its density and, hence, pressure. We should carefully note this unexpected, but nevertheless perfectly correct, conclusion. Increasing the density of gas does not lead to an increase in thermal conductivity.

Another prediction may be made on the basis of the formula for the coefficient of thermal conductivity. Since the effective cross-section and the thermal capacity hardly depend on the temperature (generally speaking,  $\sigma$  decreases slightly with increasing temperature), and the thermal velocity is proportional to  $\sqrt{T}$ , the coefficient of thermal conductivity should be proportional to the square root of the temperature.

The data presented below indicate the accuracy of these two predictions. For example, for nitrogen at 0°C, 325°C and 500°C:

$$\kappa_3 : \kappa_2 : \kappa_1 = 1.93 : 1.65 : 1,$$

$$\sqrt{T_3} : \sqrt{T_2} : \sqrt{T_1} = 1.68 : 1.48 : 1.$$

We see that the thermal conductivity increases with temperature somewhat more than proportionally to  $\sqrt{T}$ . This is due to changes in the cross-section and thermal capacity. As can be seen, the thermal conductivity is independent of pressure in a very broad interval.

Similarly, the dynamic viscosity  $\eta \sim \rho \nu l$  also does not depend on the pressure and density of the gas. The temperature dependence of the viscosity of an ideal gas should be the same as that of the thermal conductivity, i.e., the same proportionality should exist. A numerical example will help to fix this point

For nitrogen ( $T_1 = 273$  K,  $T_2 = 289$  K,  $T_3 = 296$  K):

$$\eta_3 : \eta_2 : \eta_1 = 1.06 : 1.04 : 1,$$

$$\sqrt{T_3} : \sqrt{T_2} : \sqrt{T_1} = 1.04 : 1.03 : 1.$$

---

\* It should be kept in mind that, in addition to the concept of diffusion of one substance in another, there is the concept of self-diffusion, i.e., the motion of molecules among similar molecules, e.g., the diffusion of hydrogen in hydrogen, oxygen in oxygen, etc. Investigation of this phenomenon became possible after the technique of radioactive tracers (atoms and, hence, molecules) was introduced.

Thus,  $D$  is here the coefficient of self-diffusion.

The viscosity of a gas remains amazingly constant with changing pressure. When the pressure of  $\text{CO}_2$  changes by a factor of 380—from 2 mm to 760 mm of mercury—the viscosity practically does not change. It remains at all times equal to  $14.8 \times 10^{-5}$  P—to within an accuracy of one unit in the third figure.

#### Sec. 85. ULTRA-RAREFIED GASES

When the free path length in a gas is greater than the linear dimensions of the vessel, we say the gas is ultra-rarefied. Under normal conditions, the magnitude of the free path length is of the order of  $10^{-5}$  cm and is inversely proportional to the density. Therefore, at a pressure of the order of  $10^{-4}$  mm of mercury, the free path length is measured in tens of centimetres. For vessel dimensions of about 10 cm, a vacuum or ultra-rarefied gas is obtained at such a pressure.

It should be noted that even in a vacuum the number of molecules per unit volume is large. For the pressure indicated above, 1  $\text{cm}^3$  of gas contains tens of thousands of millions of molecules.

When molecules cease to collide with one another, and collide only with the walls of the vessel, the gas acquires certain special characteristics. A number of concepts become meaningless under such conditions. Thus, it is no longer possible to speak of the internal friction of the gas molecules, since there can be no molecular layers exchanging velocities in the gas. It is no longer possible to speak of the pressure of one part of the gas with respect to another (however, the concept of gas pressure against the walls of the vessel retains its meaning). Also, the concept of heat exchange between different parts of the gas and, in general, all concepts related to interaction between different parts of the gas become meaningless. An ultra-rarefied gas interacts only with bodies within the gas.

It will be useful to illustrate by means of examples the specific character of vacuum as a special physical state of a gas.

What is the expression for the heat flow from one plate to another when these plates have different temperatures  $T_1$  and  $T_2$  and are located in a vacuum? The essence of heat exchange in this case consists in the following: gas molecules strike a wall and rebound from it with an average velocity corresponding to the temperature of this wall. As regards the expression for the heat flow, examining the familiar formula

$$q = \kappa \left| \frac{T_1 - T_2}{L} \right| = \rho c v l \left| \frac{T_1 - T_2}{L} \right|,$$

we see that the change consists in the fact that the role of free path length is now played by  $L$ , the distance between the walls. Therefore, the expression for heat flow should assume the following form in the case of ultra-rarefied gases:

$$q = \rho c v (T_1 - T_2).$$

According to this formula, when ultra-rarefied gases are rarefied still further, the heat flow should decrease after the free path length becomes comparable to the linear dimensions of the vessel. And this is precisely what experiment shows to be the case.

Also peculiar to an ultra-rarefied gas are the equilibrium conditions for a gas in two communicating vessels at different temperatures. In the case of a usual gas, the gas pressures in both vessels are the same at different temperatures, but the gas densities are different, i.e., they are inversely proportional to the temperatures. Equality of pressures is necessary for equilibrium, for otherwise gas molecules will be knocked from one vessel into the other as a result of molecular collisions.

In the case of a vacuum, the situation is completely different. In this case, no collisions occur between molecules and as a result the molecular flow between vessels is not impeded. The equilibrium condition consists in equality of molecular fluxes. If there are  $n$  particles in a unit volume and the particles move with a velocity  $v$ , then  $nv$  molecules pass through a unit area per unit time. Thus, for equilibrium,  $n_1v_1 = n_2v_2$ . Since the number of molecules in a unit volume is proportional to the pressure divided by the temperature (this follows from the equation of state for an ideal gas) and since the molecular velocity is proportional to the square root of the temperature, the condition for equilibrium assumes the form

$$\frac{p_1}{\sqrt{T_1}} = \frac{p_2}{\sqrt{T_2}}.$$

Thus, it is not the pressures that are equal, but rather the ratios of the pressures to the square root of the temperatures. If the gas density is increased, the pressures gradually begin to be equalised and the usual equilibrium condition is obtained when the free path length becomes sufficiently small.

## PART TWO

# Electromagnetic fields

### CHAPTER 14

## Electric Fields

#### Sec. 86. VECTOR PROPERTIES OF ELECTRIC FIELDS: INTENSITY AND DISPLACEMENT

The presence of an electric field in a region may be recognised by a variety of properties. Thus, an electric field creates a force that acts on electric charges. Also, it can induce electric charges on the surface of a neutral metallic body.

By measuring the force acting on a charge  $q$ , one can show that the force  $F$  has different magnitudes and directions at different points in space, and at a given point is proportional to  $q$ . Hence, it is possible to describe an electric field by its *intensity*  $E$ , which is defined as follows:

$$E = \frac{F}{q}.$$

The stipulation should be made that  $q$  must be small. Then,  $E$  may be measured at points in space that are sufficiently close to each other and the field created by charge  $q$  does not noticeably distort the measured field.

A vector field is frequently characterised by so-called vector flow lines. The tangent at each point of such a line coincides with the direction of the vector at this point. This also holds for electric fields, which may be characterised by vector flow lines of electric intensity  $E$ .

*Numerical examples.* 1. The electric field intensity of an incandescent wire is tens of volts per centimetre.

2. The electric field intensity of the Earth close to its surface is  $\sim 100 \text{ V/m} = \frac{1}{300} \text{ CGS units}$ .

3. The electric field intensity of a hydrogen atom's nucleus at a distance corresponding to the radius of the electron's "orbit" is  $19.2 \times 10^6 \text{ CGS units} = 57.6 \times 10^{10} \text{ V/m}$ .

4. The electric field intensity at which air breakdown occurs is  $30 \text{ kV/cm} = 100 \text{ CGS units}$ .

An experiment for the determination of the charge induced by a field may be conducted using two small metallic plates fastened to an insulated handle as shown in Fig. 89. Such plates are called Mie plates—after the German physicist G. Mie. Placing the pair of closely spaced plates in a field, and then carefully rotating them, positive charge may be accumulated on one plate and negative charge on the other. Moreover, the quantity of induced electricity may be measured by an electrometer or ballistic galvanometer.

Experiments show that the plates may always be so oriented that no electricity is induced on the plate faces. For homogeneous, isotropic bodies (and for the present we shall not consider others), this occurs when the plate faces are parallel to the vector  $\mathbf{E}$ . On the other hand, the induced electricity is a maximum when the plate faces are perpendicular to the vector  $\mathbf{E}$ . This enables us to introduce still another vector to describe an electric field, namely, *the electric displacement*  $\mathfrak{D}$ , which is defined by the following condition: the vector  $\mathfrak{D}$  is normal to the Mie plates when the orientation of the plates is optimal with respect to induction, i.e., when a maximum charge is induced on them. Moreover, this vector is directed outward from the positive Mie plate. In all cases, except for anisotropic bodies,  $\mathfrak{D}$  and  $\mathbf{E}$  have the same direction. The absolute value of  $\mathfrak{D}$  is equal to  $\sigma$ :

$$|\mathfrak{D}| = \sigma,$$

where  $\sigma$  is the surface charge density of the Mie plate. Since the surface charge density  $\sigma$  may be written as  $\frac{dq}{dS_{\perp}}$ , then

$$|\mathfrak{D}| = \frac{dq}{dS_{\perp}}.$$

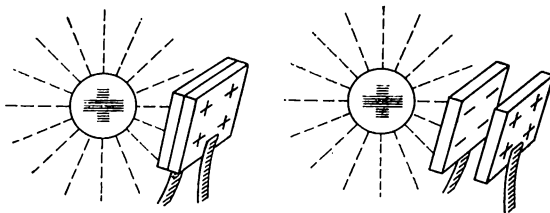


Fig. 89

It was stated above that the electric field may be characterised by the flow lines of vector  $\mathbf{E}$ . We can, of course, also describe the field by the flow lines of vector  $\mathfrak{D}$ , i.e., *the electric lines of force*. The number of lines of force passing through a unit area perpendicular to the force lines is  $|\mathfrak{D}| \equiv \mathfrak{D}$ , and the quantity

$$dN = \mathfrak{D} dS_{\perp}$$

is called *the electric flux* through the area  $dS_{\perp}$ . If the same electric flux passes through the inclined area  $dS$  as through  $dS_{\perp}$ , then

$$dS = \frac{dS_{\perp}}{\cos \alpha},$$

where  $\alpha$  is the angle between the normal to the area and the lines of force, i.e.,

$$dN = \mathfrak{D} \cos \alpha dS.$$

The flux through a large surface is expressed in the form

$$N = \int \mathfrak{D} \cos \alpha dS,$$

and the flux through a closed surface is usually denoted by placing a circle on the integral sign:

$$N = \oint \mathfrak{D} \cos \alpha dS.$$

#### sec. 87. PERMITTIVITY

Experiments show that the two vectors characterising an electric field are related. For the case when these vectors are parallel, they are also proportional to each other at a given point in space.\* A change in the vector  $\mathbf{E}$  results in a proportional change in the vector  $\mathfrak{D}$ . The ratio  $\frac{\mathfrak{D}}{E}$  depends only on the medium.

\* The case of anisotropic media, where the vectors  $\mathfrak{D}$  and  $\mathbf{E}$  are not parallel, will be considered on p. 193.

It is customary to characterise the electric properties of a medium by the dimensionless quantity  $\epsilon$ , whose value is selected so that  $\epsilon = 1$  for vacuum. The reason for this condition is the fact that, as will presently be seen, no body can exist with  $\epsilon < 1$ . Therefore, it is natural to "base" the value of  $\epsilon$  on vacuum. The quantity  $\epsilon$  is called *the permittivity* and defined by the equation

$$\frac{\mathfrak{D}}{E} = \epsilon_0 \epsilon,$$

where  $\epsilon_0$  depends on the choice of units. If the state of the medium does not vary from point to point, then  $\epsilon$  is also constant. At the boundary between two media,  $\epsilon$  changes abruptly. Bodies that are nonhomogeneous with respect to density and other properties are usually nonhomogeneous with respect to permittivity as well.

The permittivities of several substances at 18°C are as follows: air—1.00059, glass—7.00, paper—2-2.5 and water—80.5.

In the SI system, the quantity  $\mathfrak{D}$  is measured in coulombs/metre<sup>2</sup> (coul/m<sup>2</sup>), and the field intensity in newtons/coul (N/coul). Then,  $\epsilon_0$  is measured in coul<sup>2</sup>/(m<sup>2</sup> × N) and, in these units,

$$\epsilon_0 = \frac{10^{-9}}{36\pi} \frac{\text{coul}^2}{(\text{m}^2 \times \text{N})}.$$

In the CGS system,  $\epsilon_0$  is dimensionless and equal to  $\frac{1}{4\pi}$ :

$$\mathfrak{D} = \frac{\epsilon}{4\pi} E.$$

A quantity called *the electric induction*  $D$  may be used instead of displacement. It is  $4\pi$  times larger than displacement and in the CGS system  $D = \epsilon E$ .

As we shall soon see, both choices for the value of  $\epsilon_0$  have their relative advantages. The first system simplifies one group of formulas but complicates another, while the second system leads to the reverse result.

It should be emphasised that the concepts of electric displacement and electric induction have exactly the same physical meaning. The difference in the numerical factor merely leads to a difference between the ratio of an electric induction unit to a charge density unit and the ratio of a displacement unit to a charge density unit.

The electric displacement is equal to unity if the charge density on the Mie plates is equal to unity (see p. 171), while the electric induction is equal to unity if the charge density on the Mie plates is equal to  $\frac{1}{4\pi}$ .

In electrical engineering, as a rule, only the quantity  $\mathfrak{D}$ , i.e., displacement, is used. On the other hand, in physics, the electric induction  $D$  is used exclusively.

Several comments are necessary regarding the formulas and units of measurement that are used in this part of the book.

Although in mechanics and thermodynamics various choices are made for the fundamental quantities and various units of measurement are used, nevertheless, the constants of proportionality are invariably dimensionless. Therefore, the form of the formulas in those fields of physics does not depend on the choice of the system of units.

Unfortunately, however, the situation is not the same in the case of electromagnetic fields. Two general approaches exist, i.e., one approach has been adopted in electrical engineering and the other in physics. Not only is there a difference in the choice of the fundamental quantities and the units of measurement, but it turns out that we must distinguish between the constants of proportionality for one and the same formulas. Of necessity, one must become familiar with the formulas in both systems. This will be done in the course of the presentation, but for the present it suffices to limit ourselves to several general comments.

In electrical engineering, the so-called SI system is widely used. Side by side with the metre, kilogram and second, a unit of current is taken as fundamental. A current of 1 A is defined as

the current for which the constant of proportionality  $\mu_0$ , i.e., the magnetic permeability of vacuum, which occurs in formulas for electrodynamic interaction (see p. 206), has the value

$$\mu_0 = 4\pi \times 10^{-7} \text{ J/(A}^2 \times \text{m)}.$$

Experiments show that such a choice of unit current is explained by the fact that when a current of 1 A flows through two parallel conductors of an infinite length and negligibly small circular cross-section placed in vacuum at a distance of 1 m from each other, a force occurs between the conductors equal to  $2 \times 10^{-7}$  N per metre of the conductor length. The reasons for this choice, which may appear strange, will not be discussed here. For more detail see L. A. Sena "Units of Physical Quantities and Their Dimensions", MIR Publishers, 1972.

All the rest of the units in the SI system may be expressed in terms of the kilogram, metre, second, and ampere.

Since in electrical engineering the system of units is based on four fundamental quantities, there is no way of obtaining the same set of formulas in the CGS system, which is based on three fundamental quantities. There are, however, other differences between these two systems. These are expressed in the different choice of numerical, dimensionless coefficients. In the course of presentation, we shall on occasion list certain formulas in both systems, while in the appendix the reader will find a compilation of the electrodynamic formulas in both systems with the units of measurement indicated.

## Sec. 88. ELECTRIC FIELD RELATIONS

Consider a system of electrically charged bodies forming an arbitrary field. Now, describe a closed surface in this field. Some of the charges will fall within the surface, while others will be external to it. The result obtained by measuring the electric flux in the outward direction through this surface is very simple and in no way surprising. Thus, the total electric charge induced on the surface—which by definition is precisely the flux  $N = \oint \mathfrak{D} \cos \alpha dS$ —is equal to the total electric charge *within* the volume enclosed by this surface:

$$\oint \mathfrak{D} \cos \alpha dS = \sum q_i.$$

This theorem, named after Gauss and Ostrogradsky, shows that lines of electric flux begin on charges of one polarity and terminate on charges of the other polarity. Interrupted lines of force do not exist.

Lines of electric intensity closed on themselves do not exist in a constant electric field.\* This follows from the second law for electric fields, which states that an electric field (more accurately: the vector field of electric intensity  $\mathbf{E}$ ) is a potential field. Thus, the work performed in moving a charge along a closed curve is equal to zero in such a field, i.e., closed lines of vector  $\mathbf{E}$  do not exist. The work performed in moving a charge from one point to another depends only on the location of these points and does not change when the form of the path changes. In this respect, the properties of an electric field are the same as those of a gravitational field.

Let us select a reference (initial) point in an electric field and calculate potential energy with respect to this point. No matter what path is taken, the work  $A$  in moving a charge from the initial point to a given point in the field is always the same. Therefore, at this point the charge possesses a potential energy  $U$  that is numerically equal to the expended work  $A$ .

Just as the potential energy in a gravitational field is proportional to the mass of a body, the potential energy in an electric field is proportional to the charge:

$$U = \varphi q.$$

---

\* In vacuum and homogeneous media, the  $\mathbf{E}$  and  $\mathfrak{D}$  vector lines coincide. In this case, we can speak of the electric lines of force without indicating which of the vectors is being considered.

The quantity  $\varphi = \frac{U}{q}$ , i.e., the potential energy that a unit positive charge would possess at a given point in the field, is called *the electric potential of the field* or, simply, *the potential*.

The expression for the work performed in moving a charge from one point of the field to another follows from the definition of potential. Since work is equal to the change in energy, i.e.,  $dA = -dU$ , then

$$dA = F dl = qE dl = -q d\varphi,$$

$$\text{or } E dl = -d\varphi,$$

where  $d\varphi$  is the change in potential.

For a finite portion of the path

$$\int_1^2 E dl = \varphi_1 - \varphi_2.$$

Thus, the potential difference\* is equal to the work expended in moving a unit charge.

If a charge moves along a line of force, the vector sign need not be used. Then,

$$\int_1^2 E dl = \varphi_1 - \varphi_2.$$

Finally, in a uniform field, the formula is simplified to

$$E = \frac{\varphi_1 - \varphi_2}{d},$$

where  $d$  is the distance between points 1 and 2.

Formulas relating  $E$  and  $\varphi$  are written without constants of proportionality and have the same form in all systems of units.

*Examples.* 1. Assume that two flat electrodes of area  $S = 10 \text{ cm}^2$  are located in air at a distance of 5 mm apart and that the potential difference between them is 5,000 V. The intensity of the created electric field is  $E = 10^6 \text{ V/m} = 33 \text{ CGS units}$  and the electric displacement in the field of this capacitor is  $\mathfrak{D} = \epsilon_0 E = 9 \times 10^{-6} \text{ coul/m}^2$ . This means that the charge density on the capacitor plates is  $\sigma = 9 \times 10^{-6} \text{ coul/m}^2 = 2.7 \text{ CGS units}$ . The electric flux through the face of the electrode is  $N = \mathfrak{D}S = 9 \times 10^{-9} \text{ coul}$  and the charge on one plate is  $q = \sigma S = 9 \times 10^{-9} \text{ coul}$ . Thus,  $N = q$ , which agrees with the Gauss-Ostrogradsky theorem.

2. The electric displacement of the Earth's field close to its surface is  $\mathfrak{D} \sim 9 \times 10^{-10} \text{ coul/m}^2$ . Since the area of the Earth's surface is  $S \sim 5 \times 10^{14} \text{ m}^2$  and the surface charge density is  $\sigma \sim 9 \times 10^{-10} \text{ coul/m}^2$ , the Earth's charge is  $q \sim 4.5 \times 10^5 \text{ coul}$ . Thus, the electric flux passing through the Earth's surface is  $N \sim 4.5 \times 10^5 \text{ coulombs}$ .

## Sec. 89. FIELD CALCULATIONS OF SIMPLE SYSTEMS

Using the electric field relations presented in the previous article and from general considerations of symmetry, we can determine the field for certain simple systems. To determine the field means to calculate the electric intensity, induction

---

\* In a variable field, the above equation is not valid. To avoid confusion, it is convenient to introduce a separate term for  $\int_1^2 E dl$ . We refer to it as *the electromotive force (emf) along the path* between points 1 and 2. For constant fields, the emf and the potential difference are equal.



or potential. It should be noted that knowledge of the potential suffices to characterise the field. If  $\varphi$  is known at all points in space, we can determine the value of vector  $E$  by differentiating  $\varphi$ . This becomes particularly clear if we construct surfaces of equal potential (equipotential surfaces) satisfying the equation  $\varphi(x, y, z) = \text{const.}$  Since the work of moving a charge along an equipotential surface is equal to zero, the lines of force are directed normally to the equipotential surfaces. Thus, to determine the value of  $|E|$ , one must differentiate  $\varphi(x, y, z)$  in the direction of the normal. This type of mathematical operation is considered in vector analysis. However, such differentiation is easily performed graphically using a curve in which  $\varphi$  is plotted as a function of the coordinates along a line of force. The tangent of this curve's angle of inclination at any point is equal to the negative of  $E$  at that point.

In order to enable the reader to better understand these new concepts, we shall use examples to analyse the peculiarities of potential and the vector characteristics of a field. We repeat, however, that in principle knowledge of the potential suffices to solve the problem.

**Point Charge.** From considerations of symmetry, one can see that the field of a single, point charge is a radial, spherically symmetrical field.

Consider a sphere of radius  $r$ . The electric flux emanating from a charge  $q$  is equal to

$$\int \mathfrak{D} \cos \alpha \, dS = q.$$

The angle  $\alpha$  is the angle between the lines of force and the surface of the sphere, i.e., it is equal to  $90^\circ$ . At all points on the surface,  $\mathfrak{D}$  has the same value and may, therefore, be brought out in front of the integral sign. Then

$$\mathfrak{D} \oint dS = q$$

and, since  $\oint dS = 4\pi r^2$  (area of the sphere), the electric displacement at a point located at a distance  $r$  from the charge is  $\mathfrak{D} = \frac{q}{4\pi r^2}$  and the electric induction is  $D = \frac{q}{r^2}$ .

The intensity of the electric field is

$$E = \frac{q}{4\pi\epsilon_0\epsilon r^2}.$$

In this case, the CGS system of units in which  $\epsilon_0 = \frac{1}{4\pi}$  is more convenient. Then

$E = \frac{q}{\epsilon r^2}$  and in the case of vacuum

$$E = \frac{q}{r^2}.$$

Since the field intensity is equal to the derivative of the negative potential along the line of force, i.e.,

$$E = -\frac{d\varphi}{dr},$$

the expression obtained for the potential of a point charge is

$$\varphi = \frac{q}{r}.$$

The constant of integration has been assumed to be equal to zero. This fixes the reference potential  $\varphi = 0$  at infinity.

Thus, the potential of a point charge is inversely proportional to the first power of the distance, while the intensity is inversely proportional to the distance squared.

If the charge is in a medium whose dielectric constant is  $\epsilon$ , the intensity and the potential are reduced to  $\frac{1}{\epsilon}$  of the values in vacuum.

The Earth's potential is equal to 0.07 V if the potential at infinity is taken equal to zero. In electrical engineering, the potential of the Earth is assumed to be equal to zero.

**Systems of Point Charges.** Let us consider methods of calculating fields created by systems of point charges. Assume  $\epsilon = 1$  and let us use the CGS system of units. Then, the potential for a system of charges may be written in the form

$$\begin{aligned}\varphi &= \frac{q_1}{r_1} + \frac{q_2}{r_2} + \frac{q_3}{r_3} + \dots = \\ &= \sum_k \frac{q_k}{r_k},\end{aligned}$$

where  $r_k$  is the distance from the charge  $q_k$  to the point of observation.

In the case of two charges of equal magnitude, but opposite sign, we obtain

$$\varphi = q \left( \frac{1}{r_1} - \frac{1}{r_2} \right).$$

When the signs are the same:

$$\varphi = q \left( \frac{1}{r_1} + \frac{1}{r_2} \right).$$

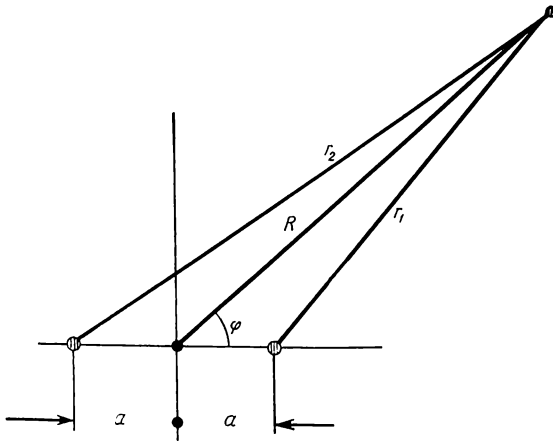


Fig. 90

The form of the above formulas may turn out to be inconvenient for solving a given problem. Thus, it is often expedient to introduce a Cartesian system of coordinates and express the radius  $r_k$  in terms of  $x$ ,  $y$ ,  $z$ . In the case of two charges, separated by a distance  $2a$ , it is convenient to locate the origin of the coordinate system at the midpoint of the system, with the  $x$ -axis passing through both charges. Then,

$$r_1^2 = (x - a)^2 + y^2 + z^2 \quad \text{and} \quad r_2^2 = (x + a)^2 + y^2 + z^2.$$

Sometimes it is expedient to represent the potential as a function of the polar coordinates  $R$  and  $\varphi$ . From Fig. 90, one can see that

$$r_1 = \sqrt{R^2 + a^2 - 2aR \cos \varphi} \quad \text{and} \quad r_2 = \sqrt{R^2 + a^2 + 2aR \cos \varphi}.$$

The field intensity of a system of point charges is given by the vector equation

$$\mathbf{E} = \frac{q}{r_1^2} \frac{\mathbf{r}_1}{r_1} + \frac{q}{r_2^2} \frac{\mathbf{r}_2}{r_2} + \frac{q}{r_3^2} \frac{\mathbf{r}_3}{r_3} + \dots = \sum \frac{q_k}{r_k^2} \frac{\mathbf{r}_k}{r_k}.$$

Here,  $\frac{\mathbf{r}_k}{r_k}$  is a unit vector in the direction of radius  $\mathbf{r}_k$ .

Vector addition is used to map the lines of force.

**Universal Formula for Potential.** When a field is created by volume and surface charges instead of point charges, the potential of the field may be calculated if the charge distribution is known.

Consider the region of the volume charge to be divided into infinitely small volumes  $dv$  and the area of the surface charge to be divided into infinitely small elements  $dS$ . If  $\rho = \frac{dq}{dv}$  is the volume density of the charge and  $\sigma = \frac{dq}{dS}$  is the surface density, the potential created by a volume  $dv$  is equal to  $\frac{\rho dv}{r}$  and the potential created by a surface element  $dS$  is equal to  $\frac{\sigma dS}{r}$ . Adding the potentials created by all the elements, we obtain:

$$\varphi = \int \frac{\rho dv}{r} + \int \frac{\sigma dS}{r}.$$

The radius  $r$  is drawn from the point of observation to all the points in space where charges  $\rho dv$  and  $\sigma dS$  are concentrated.

This formula is rarely used since the charge distribution as a function of  $\rho$  and  $\sigma$  is not usually given. In fact, the charge distribution is generally being sought.

**Field of a Spherical Condenser.** Consider a sphere of radius  $r_A$  having a charge  $+q$  and surrounded by a concentric, spherical surface of radius  $r_B$ . It is convenient to view the external sphere as grounded, whereupon a charge  $-q$  is induced on its inner surface. Considerations of symmetry indicate that the field is radial. If we describe a sphere of radius  $r$  between the condenser (capacitor) spheres and apply the Gauss-Ostrogradsky theorem, the result obtained does not differ from that for a point charge, i.e.,

$$E = \frac{q}{r^2}.$$

The potential equation has the form

$$\varphi = \frac{q}{r} + \text{const},$$

but the constant in this case should not be discarded as was done previously. As is well known, the potential of grounded metallic parts is usually taken equal to zero. It will be, therefore, more convenient to set  $\varphi = 0$  at  $r = r_B$  rather than at infinity. We then obtain:  $\text{const} = -\frac{q}{r_B}$ .

The expression for the potential in the region between the spheres has the form

$$\varphi = \frac{q}{r} - \frac{q}{r_B}.$$

On the surface of the inner sphere,

$$\varphi = \frac{q}{r_A} - \frac{q}{r_B}.$$

Recalling that the ratio of the charge to the potential difference between the conductors (or plates) of a condenser gives the *capacitance*, we obtain for the capacitance of a spherical condenser

$$C = \frac{1}{\frac{1}{r_A} - \frac{1}{r_B}} = \frac{r_A r_B}{r_B - r_A}.$$

If the radius of the outer sphere is increased ( $r_B \rightarrow \infty$ ), the expression for the capacitance is reduced to

$$C = r.$$

Thus, the capacitance of a single sphere is equal to the magnitude of its radius.

If the dielectric between the conductors of the condenser has a permittivity  $\epsilon$ , the intensity  $E$  and the potential  $\varphi$  decrease to  $\frac{1}{\epsilon}$  of the above values.

From the formula

$$\varphi = \frac{1}{\epsilon} \left( \frac{q}{r_A} - \frac{q}{r_B} \right),$$

we obtain for the capacitance of the condenser:

$$C = \epsilon \frac{r_A r_B}{r_B - r_A};$$

and for a sphere,

$$C = \epsilon r.$$

Thus, the capacitance is  $\epsilon$  times the value obtained for vacuum.

The potential and field formulas being used are applicable for points in the region between the conductors of a condenser. They are not applicable to points

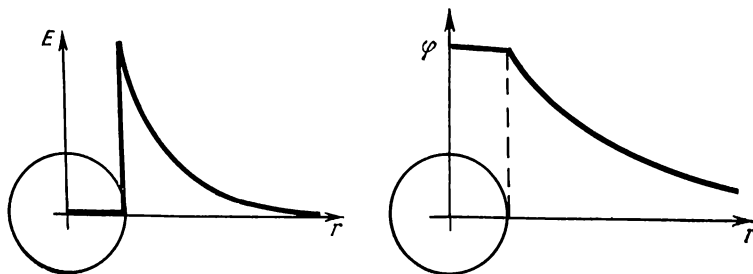


Fig. 91

within the first conductor or external to both conductors, for Gauss's theorem yields different results for these points.

If the charge of the internal sphere is concentrated on its surface, then for points within the sphere

$$\oint D \cos \alpha dS = 0.$$

Since an analogous relation is valid for every surface within the sphere, this requires that  $D = 0$  and, hence, the field intensity is also equal to zero. Thus, Gauss's theorem shows that there is no field within the sphere if all the charge is distributed on its surface. Since  $E = 0$ , the potential  $\varphi$  remains constant and equal to the value of  $\varphi$  on the surface of the sphere. The above is illustrated in Fig. 91 by the curves of  $E$  and  $\varphi$  as functions of  $r$ .

*Examples.* 1. The electric field intensity at the Earth's surface is

$$\left( \frac{7,400}{6,400} \right)^2 = 1.33$$

times the intensity at a distance of 1,000 km from the surface.

2. The Earth's capacitance is  $C = 6.4 \times 10^8$  CGS units = 700  $\mu\text{F}$ .

3. The capacitance of condensers used in radio engineering may be as small as a fraction of a picofarad (1 pF =  $10^{-12}$  F) and as large as thousands of microfarads.

**Field of a Uniformly Charged Sphere.** It is evident that outside such a sphere the field is the same as for a point charge or a surface-charged sphere, i.e.,

$$E = \frac{q}{r^2},$$

where  $r$  is the distance from the centre of the sphere and  $q = \frac{4}{3}\pi a^3\rho$  ( $\rho$  is the charge density and  $a$  is the radius of the sphere).

To determine the field inside the sphere, consider an auxiliary sphere of radius  $r < a$ . The quantity of electricity inside this sphere is less than  $q$ , being equal to

$$\frac{4}{3}\pi r^3\rho = \frac{r^3}{a^3}q.$$

According to Gauss's theorem

$$\mathfrak{D} \times 4\pi r^2 = \frac{r^3}{a^3}q,$$

i.e.,

$$\mathfrak{D} = \frac{q}{4\pi a^3}r.$$

Hence, the electric field intensity is

$$E = \frac{q}{4\pi\epsilon_0\epsilon a^3}r \text{ (SI)}$$

or

$$E = \frac{q}{\epsilon a^3}r \text{ (CGS)}.$$

It should be noted that the field is equal to zero only at the centre of the sphere.

Then, as shown in Fig. 92, the field increases linearly and becomes equal to  $\frac{q}{\epsilon a^2}$

at the surface of the sphere ( $r = a$ ). Here, the formula for the field outside the sphere and the formula for the field inside the sphere yield the same result. From this radius outward, the field decreases in accordance with an inverse square relationship. The potential outside such a sphere is again given by  $\frac{q}{r}$ . Inside the sphere, the value of  $\varphi$  does not interest us and will not be considered.

**Cylindrically Radial Field.** Let us consider the field created by a uniformly charged line or cylinder having a charge  $\frac{q}{l}$  per unit length. Outside the

charged region, the fields of such systems are the same and have the following form: the lines of force are at right angles to the axis of symmetry and the flux is the same in all radial directions.

In order to apply Gauss's theorem, let us consider an auxiliary cylindrical surface of radius  $r$  and unit height. Since the flux passes only through the lateral surface of this cylinder,  $\oint \mathfrak{D} \cos \alpha dS$  is equal to the integral over the lateral surface. Owing to symmetry ( $\cos \alpha = 1$  and  $\mathfrak{D}$  is the same at all points of the cylinder),

$$\mathfrak{D} \int dS = \frac{q}{l},$$

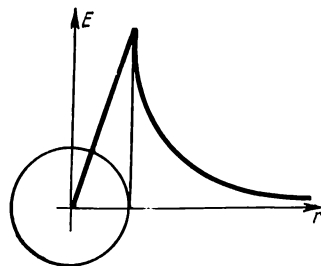


Fig. 92

i.e.,

$$\mathfrak{D} \times 2\pi r = \frac{q}{l} \quad \text{and} \quad \mathfrak{D} = \frac{q}{2\pi r}.$$

Hence, the field intensity is

$$E = \frac{q}{2\pi\epsilon_0\epsilon r} \quad (\text{SI})$$

or

$$E = \frac{2}{\epsilon r} \frac{q}{l} \quad (\text{CGS}).$$

Thus, the field of the cylinder is inversely proportional to the distance. This formula is equally valid for the region about a charged line, the region outside a charged cylinder and the region between the conductors of a cylindrical condenser.

Since  $d\varphi = -E dr$ , we obtain for the potential:

$$\varphi = \frac{2}{\epsilon} \frac{q}{l} \ln \frac{1}{r} + \text{const.}$$

The potential decreases much slower with increasing distance than in the case of spherical systems. Thus, for example, when the distance  $r$  is increased 10 times its original value, the potential decreases to  $\frac{1}{2.3}$  of its value rather than to  $\frac{1}{10}$ .

For a cylindrical condenser, where the radii of the cylinders are  $a$  and  $b$ , we obtain

$$\varphi_b - \varphi_a = \frac{2}{\epsilon} \frac{q}{l} (\ln a - \ln b) = \frac{2}{\epsilon} \frac{q}{l} \ln \frac{a}{b}.$$

The capacitance per unit length of such a condenser is

$$C = \frac{\epsilon}{2 \ln \frac{a}{b}}.$$

*Example.* For a coaxial cable having an outer radius  $a = 18$  mm and an inner radius  $b = 6$  mm, and filled with an insulator of relative permittivity  $\epsilon = 4.2$ , the capacitance per unit length is  $C = 4.91$  CGS units/cm  $= 2.42$  pF/cm.

It should be noted that the formulas derived above do not take account of field distortion at the ends of the cylinder and hence, strictly speaking, are valid only for infinitely long cylinders. Practically, however, the derived formulas are valid if the region of "distorted" field is significantly smaller than the undistorted radial field.

**Uniform Fields.** Uniform fields, i.e., fields in which the lines of force are parallel and evenly spaced, are created by charges in planes of infinite extent. Naturally, the flux is perpendicular to such planes. The magnitude of the field is again determined by means of the Gauss-Ostrogradsky theorem. Thus, let us consider an auxiliary surface in the form of a cylinder passing through a charged plane. If the lateral surface of the cylinder is perpendicular to the plane, the flux through the auxiliary surface is equal to the flux through the two end surfaces of the cylinder. The integral  $\oint \mathfrak{D} \cos \alpha dS$  is then equal to  $2\mathfrak{D}S$ , where  $S$  is the area of the cylinder base. The charge within the cylinder is equal to  $\sigma S$ . Hence, the formula

for the displacement becomes

$$\mathfrak{D} = \frac{\sigma}{2}.$$

The electric field intensity is  $E = \frac{\sigma}{2\epsilon\epsilon_0}$  in the SI system and in the CGS system  $E = \frac{2\pi\sigma}{\epsilon}$ . We see that the intensity does not depend on the distance to the sources of the field. Let us now consider a parallel-plate condenser. The field inside spherical and cylindrical condensers is created only by the inner charged surface while in a parallel-plate condenser the field between the plates is created by both surfaces. Just as in the case of the condensers considered above, there will be no field outside the condenser. Between the condenser plates  $\mathfrak{D} = \sigma$  and the intensity is

$$E = \frac{4\pi}{\epsilon} \sigma \text{ (CGS).}$$

In writing the expression for the potential of a uniform field, let us reckon the distance  $x$  from one of the charged plates in the direction of the lines of force. Thus, in the case of a single plate, this potential is written in the form  $\varphi = -\frac{2\pi}{\epsilon}\sigma x + \text{const.}$  In the case of a condenser, the expression for the potential between the plates becomes

$$\varphi = -\frac{4\pi}{\epsilon}\sigma x + \text{const.}$$

Hence, the potential difference is

$$\varphi_a - \varphi_b = \frac{4\pi}{\epsilon}\sigma(x_b - x_a) = \frac{4\pi}{\epsilon}\sigma d,$$

where  $d$  is the distance between the plates. Thus, the capacitance per unit area of a parallel-plate condenser is

$$C = \frac{\epsilon}{4\pi d} \text{ (CGS)} \quad \text{or} \quad C = \frac{\epsilon_0\epsilon}{d} \text{ (SI).}$$

The above formulas are exact only for plates of infinite extent. In practice, they may be employed if the effect of the condenser edges, where the nonuniformity of the field is pronounced, is not great. We can determine the field at some point by means of the derived formulas only if this point is sufficiently far from the edges. More specifically, this condition means that the field created by elementary charges located at the edges of the plates should be much less than the field created in the neighbourhood of the point under consideration.

*Example.* Let us return to the condenser considered on p. 175. The distance between the plates will be doubled using two different methods:

Method 1: The plates remain connected to a source of voltage  $U = 5 \text{ kV}$ . Then,  $C_1 = \frac{\epsilon_0 S}{d_1} = 1.8 \text{ pF}$ ,  $q_1 = C_1 U = 9 \times 10^{-9} \text{ coul}$ ,  $E_1 = \frac{U}{d_1} = 10^6 \text{ V/m}$ ,  $\mathfrak{D}_1 = 9 \times 10^{-6} \text{ coul/m}^2$  and  $N_1 = 9 \times 10^{-9} \text{ coul}$ .

After doubling the distance between the plates, we obtain:  $C_2 = 0.9 \text{ pF}$ ,  $q_2 = 4.5 \times 10^{-9} \text{ coul}$ ,  $E_2 = 0.5 \times 10^6 \text{ V/m}$ ,  $\mathfrak{D}_2 = 4.5 \times 10^{-6} \text{ coul/m}^2$  and  $N_2 = 4.5 \times 10^{-9} \text{ coul}$ . Thus, half the charge entered the source.

Method 2: Before doubling the distance, let us disconnect the plates from the source (condenser charge  $q = \text{const.}$ ).  $C_2 = 0.9 \text{ pF}$ ,  $q = q_1 = 9 \times 10^{-9} \text{ coul}$ ,  $U_2 = \frac{q}{C_2} = 10 \text{ kV}$ ,  $E_2 =$

$= \frac{U_2}{d_2} = E_1$ ,  $\mathfrak{D}_2 = \mathfrak{D}_1$ , and  $N_2 = N_1$ . Thus, the voltage on the plates doubled at the expense of the work of external forces.

**Field on the Surface of a Metal Object.** There is no electric field inside a metal object. This follows from the fact that all of a conductor's charge is located on its surface. According to Gauss's theorem, the field is directed outwardly.

The surface of a metal object is obviously an equipotential surface, for otherwise the electric charges would redistribute themselves on the surface of the conductor. It follows, therefore, that the lines of flux leaving the surface of the metal object

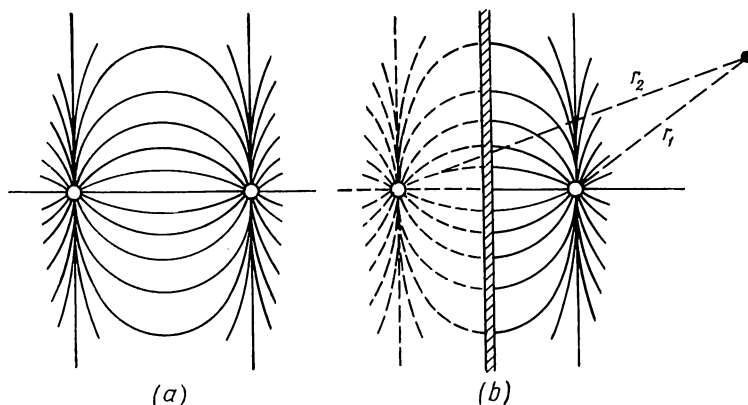


Fig. 93

must be perpendicular to the surface. Since all the flux leaves the surface in a single direction, then, according to Gauss's theorem,  $\mathfrak{D} = \sigma$  lines emerge from unit surface area. In other words, the field intensity at the surface of the conductor is equal to  $E = \frac{4\pi\sigma}{\epsilon}$  (CGS).

**Electric Images.** Let us consider the electric field created when a point charge is placed near a plane metallic surface. Due to electric induction, an electric charge of opposite sign accumulates on the surface of the metal near the point source. The density of the induced charge is greatest directly opposite the point source and decreases to zero at infinity. Similarly, for the electric field.

Let us now consider this problem quantitatively. Since the surface of a conductor is an equipotential, the conducting surface may be considered grounded without in any way reducing the generality of the problem. Hence, the potential of the metal plate is equal to zero and there is no field inside the plate. We are interested in the electric field in the right half-space. The electrical properties of this half-space are uniquely determined if the magnitude of the charge and its distance from the equipotential plane are given. It is important to note that it is entirely immaterial what is located to the left of the zero-potential surface. It is proved rigorously in courses on mathematical physics that the field in a particular region is uniquely determined if the charge in this region and the boundary conditions for the potential are given.

In Fig. 93a, the field is plotted for two charges of opposite sign. Now, consider the space of this field to be divided into two symmetrical parts. The half-space of this figure is then exactly equivalent to the half-space of a charge near a metal plate



(Fig. 93b) and the fields of such half-spaces should be identical. This is the basis for the following procedure, known as the method of images. We “reflect” the electric charge in the surface of the metal plate. In the right half-space, the electric field of the charge and its “image” should coincide with the unknown field. Thus, the unknown electric field is expressed by the formula

$$\varphi = q \left( \frac{1}{r_1} - \frac{1}{r_2} \right),$$

where  $r_1$  is the distance of the observed point from the charge and  $r_2$  is the distance of this point from the charge's image.

The second conclusion to be drawn is that the electric charge is attracted to the surface of the metal plate with the same force as to its electric image, i.e., the force of attraction is  $\frac{q^2}{4a^2}$ , where  $a$  is the distance of the charge from the surface.

Finally, this approach to the problem enables us to determine the distribution of the induced electric charge on the surface of the metal plate. This requires that we differentiate the expression for the potential in the direction normal to the surface. We obtain thereby the electric field intensity  $E$ , which in accordance with the formula given in the preceding section of this article must be multiplied by  $\frac{\epsilon}{4\pi}$  to obtain the charge.

The method of electric images has numerous applications and enables us to solve electrostatic problems involving systems of nonplanar metal conductors with point charges located in the vicinity of the conductors.

#### Sec. 90. ELECTRIC ENERGY

**Energy of a Condenser.** It is easily demonstrated that a charged electric condenser possesses energy. Moreover, the measurement of the magnitude of this energy is not difficult. Thus, for example, we can discharge a condenser through a conductor and measure the Joule heat released thereby. However, we need not resort to experiment to determine the factors on which the electric energy of a condenser depends. The formula for this energy follows directly from familiar theoretical propositions.

To simplify this discussion, let us consider a condenser in which one of the conductors is grounded. The process of discharging the condenser (grounding the second conductor), which is charged to a potential difference  $\varphi$  by a quantity of electricity  $q$ , may be viewed as the successive outflow to ground of elementary charges  $dq$  under the action of electric field forces. Therefore, the work performed by the field in this simple process is equal to  $\varphi dq$ . As the discharging process proceeds, the work performed in transferring each successive quantity of charge to ground becomes less and less, for the potential difference  $\varphi = \frac{q}{C}$  is constantly decreasing. The total work performed by the field during condenser discharge is

$$\int_0^q \frac{q}{C} dq = \frac{q^2}{2C}.$$

This quantity is quite understandably referred to as the *electric energy of the condenser*. Using the relationship between potential and charge, we obtain for the energy:

$$W_{el} = \frac{q^2}{2C} = \frac{\varphi q}{2} = \frac{C\varphi^2}{2}.$$

Thus, for constant potential difference, the electric energy is proportional to the charge squared. A constant difference of potential is maintained when the condenser is connected to a constant source. Moreover, if the condenser conductors are insulated, the charge is constant. Then, the energy is proportional to the potential squared and directly proportional to the capacitance of the condenser.

**Energy of a Field.** In the case of a parallel-plate condenser of infinite extent, the energy formula may be written in the form  $W_{el} = \frac{\epsilon_0 \epsilon q^2}{2d}$  when using the SI system<sup>1</sup> or in the form  $W_{el} = \frac{\epsilon q^2}{8\pi d}$  when using the CGS system. These formulas give the energy per unit condenser area.

The energy formulas may be written in terms of intensity  $E$  rather than potential difference  $\varphi$ . Making the substitution  $\varphi = Ed$ , we obtain

$$W_{el} = \frac{\epsilon_0 \epsilon E^2}{2} d \quad (\text{SI}), \quad \text{or} \quad W_{el} = \frac{\epsilon E^2}{8\pi} d \quad (\text{CGS}).$$

Thus, the energy per unit volume is  $\frac{\epsilon E^2}{8\pi}$ . Let us call  $w = \frac{\epsilon E^2}{8\pi}$  the *electric energy density*.

Now, let us consider an arbitrary electric field. Assume that the equipotential surfaces and lines of force are plotted and that the space is divided into small volumes  $dv$ , each of which is bounded by two adjacent equipotential surfaces and a lateral surface passing through lines of force. Each of these volumes is like a small volume in a parallel-plate condenser and, hence, the electric energy associated with such an element is  $dW = \frac{\epsilon E^2}{8\pi} dv$ . If this expression is integrated over the entire volume occupied by the electric field, the formula obtained yields the electric energy of the system creating the field.

Thus, the formula for the electric energy has the form

$$W_{el} = \int \frac{\epsilon E^2}{8\pi} dv.$$

The significance of the above mathematical transformations goes beyond the formal convenience of using one or another formula. This new expression for the energy enables us to speak not only of the energy of the system creating the field, but of the energy of the electric field itself, and leads to the concept of a real electric field. In the case of constant fields, this conception can neither be confirmed nor refuted. However, in the case of varying fields, we find direct evidence for the existence of electromagnetic fields (see p. 242). Hence, the derived formula for the energy of a field (energy of electromagnetic matter) is of fundamental significance.

*Example.* Let us continue with the example considered on p. 181. Before the plates were moved, the energy stored in the electric field of the condenser was  $W_1 = 22.5 \times 10^{-6}$  J and the energy density  $w_1 = 4.5$  J/m<sup>3</sup>. After the plates are moved by the first method (voltage  $U = \text{const}$ ), the energy becomes  $W_2 = \frac{W_1}{2} = 11.25 \times 10^{-6}$  J and the energy density  $w_2 = 1.12$  J/m<sup>3</sup> (the volume of the field doubled). The energy of the source increases at the expense of the work of external forces and a decrease in the energy of the field. After the plates are moved by the second method ( $q = \text{const}$ ), the energy  $W_2 = 2W_1 = 45 \times 10^{-6}$  J and the energy density does not change, i.e.,  $w_2 = 4.5$  J/m<sup>3</sup>.

**Energy of Interaction.** When two oppositely charged bodies draw together, the work performed by the forces of the electric field is, naturally, at the expense of the energy of the electric field:  $dA = -dW_{el}$ . Thus, as indicated on p. 39,

the work of the electric forces is performed at the expense of a decrease in the potential energy  $\frac{q_1 q_2}{r}$ . This energy is appropriately called *the interaction energy of the charges*.

What is the relation of this formula to the formula for the electric field energy considered above? It is evident that the interaction energy is a part of the electric field energy of the charges under consideration. Carefully examining the formula for the field energy, we note that the electric energy has definite meaning even when there is only one electric charge in the region. The energy of the field created by a single charged body is appropriately called *the self-energy of the electric charge*. We can always resolve the electric field energy into the self-energies of the individual electric charges and the interaction energies of these charges.

Let us designate by  $E_1, E_2, \dots$  the field intensity created by the first, second, etc., charges. The total field is equal at each point in space to the vector sum of the intensities:  $E = E_1 + E_2 + \dots$ .

The electric energy density is

$$\frac{\epsilon}{8\pi} (E_1 + E_2 + \dots)^2 = \frac{\epsilon}{8\pi} E_1^2 + \frac{\epsilon}{8\pi} E_2^2 + \dots + \frac{\epsilon}{4\pi} E_1 E_2 + \frac{\epsilon}{4\pi} E_1 E_3 + \dots$$

Clearly, the individual terms of this expansion correspond to the energy components discussed above. Thus, the  $E$ -squared terms yield the self-energies and the terms involving the product of two different intensities yield the interaction energies. The interaction energies of the charges may be positive or negative quantities. On the other hand, the self-energies of the charges and the total energy of the field must be positive.

As a rule, only the interaction energies of electric charges are involved in a problem. We can calculate, therefore, the work of electric forces by determining the decrease in the energy of interaction or in the energy of the field. The easier calculation is the one that should be performed.

#### Sec. 91. ELECTRON RADIUS AND THE LIMITATIONS OF CLASSICAL ELECTRODYNAMICS

Let us calculate the self-energy of a spherical charge, assuming the electricity to be distributed on the surface. The electric field is then only outside the charge. Therefore, the energy of the field must be integrated over the region external to the sphere. If the charge is in a vacuum, the field intensity is expressed by the formula  $\frac{q}{r^2}$  and the energy density at any point in the region has the form  $\frac{1}{8\pi} \frac{q^2}{r^4}$ . Consider the entire region to be divided into spherical shells. The energy contained in such a shell, whose inner radius is  $r$  and outer radius  $r + dr$ , is  $\frac{1}{8\pi} \frac{q^2}{r^4} \times \text{vol. of shell}$ . Since the volume of the shell is equal to  $4\pi r^2 dr$ , the energy in this spherical shell is given by the simple expression  $\frac{1}{2} \frac{q^2}{r^2} dr$ . To determine the total energy of the field, this expression must be integrated from  $a$  (radius of the spherical charge) to infinity. Thus,

$$W = \frac{q^2}{2} \int_a^\infty \frac{dr}{r^2} = \frac{q^2}{2a}.$$

This is the form of the energy formula for an electrically charged sphere.

We shall leave it to the reader to show that if the charge is distributed throughout the volume of the sphere the energy formula obtained is almost the same. The only difference is that in this case the formula contains a coefficient close to unity.

What is the result if the above formula is applied to an elementary particle, e.g., an electron?

According to the principle of relativity (see p. 320), the internal energy of a body of mass  $m$  is given by the expression  $mc^2$ , where  $c$  is a universal constant equal to the propagation velocity of electromagnetic waves in vacuum. Equating the two energy expressions, we obtain the formula for the electron radius:

$$a = \frac{q^2}{2mc^2}.$$

Substituting numerical values\* in this interesting formula, we obtain  $a = 1.4 \times 10^{-13}$  cm. There is considerable indirect evidence in physics that the order of magnitude of the electron radius determined in this manner is quite correct.

Nevertheless, the conception of an electron as a "usual" electric particle is clearly false, for we are immediately confronted with the problem of the forces holding the component parts of an electron so close together. We know that the forces of repulsion between electric particles separated by a distance of the order of  $10^{-13}$  cm are tremendous.

Furthermore, there are other theoretical difficulties. Thus, it follows from the theory of relativity that an electron should be a mathematical point. At the same time, the electric energy of a charge concentrated at a point is infinitely great.

These difficulties are typical for so-called classical physics, which developed in the main in the 19th century. Classical physics excellently describes the behaviour of macroscopic bodies. In fact, by the turn of the century many scientists believed that classical physics was already so perfected that there was little left to be discovered in physics. After the discovery of elementary particles, it was natural to try to apply the laws established for large bodies to elementary particles. This is when classical physics began to "fail". We now know that concepts derived from observations on macroscopic systems cannot be simply transferred to atoms, nuclei and electrons.

The electron problem cannot be solved within the framework of classical conceptions. Considerable success has been achieved in electron theory during recent years, but a complete theory does not exist. Therefore, the classical theory of electricity (electrodynamics) presented in this part of the book has certain limitations. These are encountered in studying the interaction of elementary particles. When dealing with the behaviour of a single elementary particle in fields created by large bodies and, of course, when considering the interaction of macroscopic bodies, one obtains, using classical electrodynamics, results that are in complete agreement with experimental data.

## Sec. 92. ELECTRIC FORCES

In calculating interaction forces between charged bodies, one often uses the concept of an electric force field. Instead of saying that body A exerts a certain force on body B, we introduce a force field and say that body A creates a field and this field acts on body B. As we shall see in Chapter XVI, this field is more than

---

\*  $q = 1.602 \times 10^{-19}$  C;  $m = 9.1091 \times 10^{-31}$  kg; and  $c = 2.99792 \times 10^8 \frac{\text{m}}{\text{sec}}$ .

an abstraction. An electromagnetic field is a physical reality and nature implements the interaction transmitted from one point in space to another ("short range" action). By introducing the field concept, we can ignore the field sources and determine the forces acting on a charged body knowing only the field intensities where the charges of the system under consideration are located.

Every charged body is a system of charges. In the case of a system of discrete charges, the force acting on the system is  $\mathbf{F} = q_1\mathbf{E}_1 + q_2\mathbf{E}_2 + \dots$ . Here  $\mathbf{E}_1, \mathbf{E}_2, \dots$  are the field intensities where the charges are located. When the electric charge is uniformly distributed throughout a volume, the force acting on the body may be represented by the following integral:  $\mathbf{F} = \int E\rho \, dv$ . If the electric charge is distributed over a surface, the force is represented by a surface integral:  $\mathbf{F} = \int \mathbf{E}\sigma \, dS$ .

However, one precaution must be taken when the force is determined directly in this manner, namely, the value of the intensity used in the formulas must be the intensity existing in the absence of the charge on which the force is acting. In the formulas in which the force is expressed as a sum, the action of charge  $q_i$  on itself does not enter into intensity  $\mathbf{E}_i$ , i.e., in calculating  $\mathbf{E}_1$  the field created by  $q_1$  is not considered, etc. The same is true for the integral formulas, i.e., the field intensity under the integral sign is the intensity created by the entire distribution of electric charge, except for the quantity of electricity located at the point under consideration.

Let us illustrate this by means of the force acting on a charged metal surface. As we know, the electric field intensity on the surface of a metal bounded by a dielectric is equal on the dielectric side to  $\frac{4\pi\sigma}{\epsilon}$  in the CGS system, and on the metal side is equal to zero. The field intensity is broken on this surface. To determine the force acting on an element of surface, we must multiply the quantity of electricity  $\sigma \, dS$  by the intensity which would exist at this location if the element of charged surface under consideration were removed. Therefore, it would not be correct to multiply  $\sigma \, dS$  either by  $\frac{4\pi\sigma}{\epsilon}$ , the value of the field on the dielectric side, or by zero, the value of the field on the metal side. It can be shown rigorously that the field existing at this location after removing the element under consideration is equal to the arithmetic mean of 0 and  $\frac{4\pi\sigma}{\epsilon}$ , i.e., is equal to  $\frac{2\pi\sigma}{\epsilon}$ . Thus, the formula for the force acting on an element of a conducting body's charged surface has the form

$$\frac{2\pi\sigma^2}{\epsilon} dS,$$

and for the entire body

$$F = 2\pi \int \frac{\sigma^2}{\epsilon} dS.$$

The integration in the above formula must be performed over the entire surface, taking into consideration differences in charge density and dielectric constant along the metal surface.

In the case of a uniform field (ideally, between the plates of a condenser of infinite extent), the force  $F$  acting on the plate area  $S$  may be determined with considerable accuracy by the formula

$$F = \frac{2\pi\sigma^2}{\epsilon} S \text{ (CGS).}$$

The magnitude of this force may be measured by means of a Thomson balance, whose method of operation is illustrated in Fig. 94. When the potential difference between condenser plates of 50 cm<sup>2</sup> area is 600 V and the distance between plates is 5 mm, the force of attraction between the plates, calculated in the two systems of units that we have been using, may be determined as follows:

*The CGS system*

$$C = \frac{eS}{4\pi d} = 8 \text{ cm}$$

$$q = CU = 8 \times 2 = 16 \text{ CGS units}$$

$$\sigma = 0.32 \text{ CGS units}$$

$$F = \frac{2\pi\sigma^2}{\epsilon} S = \frac{2 \times 3.14 \times 0.32^2}{1} \times 50 = 32 \text{ dynes}$$

*The SI system*

$$C = \frac{\epsilon\epsilon_0 S}{d} = \frac{1 \times 10^{-9}}{36\pi} \times \frac{5 \times 10^{-3}}{5 \times 10^{-3}} = 8.9 \times 10^{-12} \text{ F}$$

$$q = CU = 8.9 \times 10^{-12} \times 600 = 5.3 \times 10^{-9} \text{ coul}$$

$$\sigma = 1.06 \times 10^{-6} \text{ coul/m}^2$$

$$F = \frac{\sigma^2}{2\epsilon\epsilon_0} S = \frac{(1.06 \times 10^{-6})^2 \times 5 \times 10^{-3}}{2 \times 1 \left[ \frac{10^{-9}}{36\pi} \right]} = 32 \times 10^{-5} \text{ N}$$

Thus, in order to balance the force of electrostatic attraction, one must place in the opposite pan a weight equal to 32 dynes =  $32 \times 10^{-5}$  N.

It is still more difficult to determine the force acting on a body having a distribution of volume charge. Here, in the expression  $\rho E dv$ , the intensity  $E$  is the intensity of the field created by all the charges except  $\rho dv$ .

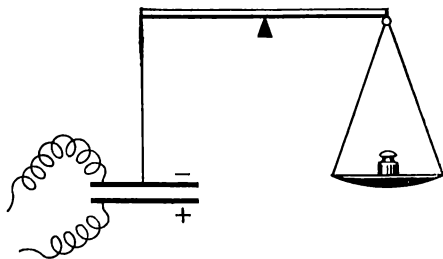


Fig. 94

If the charged body is in a dielectric medium, calculation of the force is complicated by the fact that when we consider the charge to be removed it is also necessary to consider a corresponding portion of the dielectric removed, which means the polarised state changes (see below).

If we wish to avoid the difficulties connected with "subtracting" the effect of the charge on itself, the force must be determined from the energy expression. The decrease in energy is equal to the work. Then, if we know the magnitude of the displacement, we can determine the value of the force. As a rule, this is precisely the method used in determining the force.

Calculation by this method of the force acting on the plate of a parallel-plate condenser,  $F = \frac{2\pi\sigma^2}{\epsilon} S$ , may serve as a vivid illustration of the above. Observing the attraction between the plates of the condenser (disconnected from a source of voltage), we can immediately write the expression for the change in energy when the plates come together by an amount  $\Delta$ :

$$S\Delta \frac{\epsilon E^2}{8\pi} = \frac{2\pi\sigma^2}{\epsilon} S\Delta.$$

Hence, the force sought is

$$F = \frac{2\pi\sigma^2}{\epsilon} S.$$

## Sec. 93. DIPOLE MOMENT OF A SYSTEM OF CHARGES

Let us return to electrical systems that may be represented as systems of point charges. Assume that the electric field is uniform in the region of the system of charges under consideration. Then, the formula for the force acting on the system has the form

$$\mathbf{F} = (q_1 + q_2 + \dots) \mathbf{E} = Q\mathbf{E},$$

where  $Q$  is the total charge of the system. If a body is electrically neutral, as in the case of an atom or a molecule, the force acting on the body, which contains equal quantities of positive and negative particles, is equal to zero. Does this mean that an electrically neutral body does not interact with the electric field? It is easily seen that the answer is no. In a uniform field, the forces acting on the charges of the system are parallel to each other. We can determine the resultant of the forces acting on the positive charges and the resultant of the forces acting on the negative charges. As is well known, the resultant of parallel forces is exerted at the centre of "gravity" of a body. Since we are dealing here with the electric centre of gravity, the word "gravity" has been placed in quotation marks. Thus, all the forces acting on the charges of a system located in a uniform field may be reduced to two antiparallel forces. One of the forces is applied at the centre of gravity of the positive charges and the other at the centre of gravity of the negative charges (Fig. 95). If the system is electrically neutral, the two forces are equal and the total force is zero. However, a couple of forces of moment  $M = q\ell l \sin \alpha$  will act on the system of charges if the centres of "gravity" of the positive and negative charges are displaced with respect to each other.

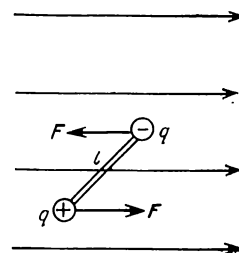


Fig. 95

The vector  $\mathbf{p} = q\mathbf{l}$ , equal in magnitude to the product of the positive charge of the system and the distance between the centres of gravity, is called the *dipole moment* of the system. It is considered to be directed from the negative centre to the positive centre. The dipole moment of a system determines its behaviour in a uniform field. Thus, a system left to itself in a uniform electric field tends to turn until the dipole moment is parallel to the direction of the electric field ( $\sin \alpha = 0$ ).

In a uniform field, the entire effect on a neutral system of electric charges is reduced to the moment of force  $M = pE \sin \alpha$ , where  $p$  is the dipole moment of the system and is equal to the product of the quantity of electricity of one sign and the distance between the dipole charges. Thus, in a uniform field, there is no need to consider the complex distribution of a particular system of charges. It suffices to replace it by the corresponding dipole.

If the system is located in a nonuniform field, the dipole moment can no longer completely describe its behaviour. This is illustrated in Fig. 96 where four charges, located at the corners of a square, comprise an electrically neutral system having a dipole moment equal to zero. This is because the centres of gravity of the negative and positive charges coincide. In a uniform field, neither a force nor a moment of force acts on such a system. In a nonuniform field, however, it may undergo translatory as well as rotational motion, since generally speaking the forces acting on the charges differ. By analogy with the dipole, such a system of four charges is called a *quadrupole*. Another neutral system having zero dipole moment, called an *octupole*, is also shown in Fig. 96.

Of great importance in the study of the structure of matter, to which we shall devote a great deal of attention later, is the consideration of the interaction of simple electric systems. Let us consider several such systems.

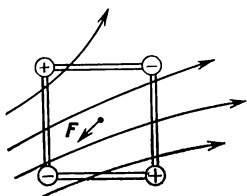


Fig. 96

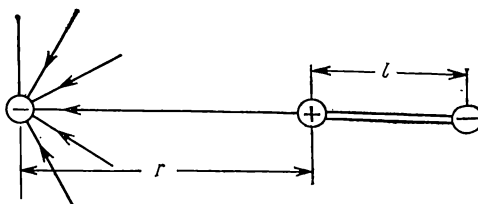
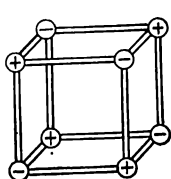


Fig. 97

**Charge-Charge.** The interaction between two point charges is, in accordance with Coulomb's law,  $F = \frac{q_1 q_2}{r^2}$ .

**Charge-Dipole.** A dipole left to itself tends to turn until it is parallel to the lines of force. After it has turned, the dipole remains stationary in a uniform

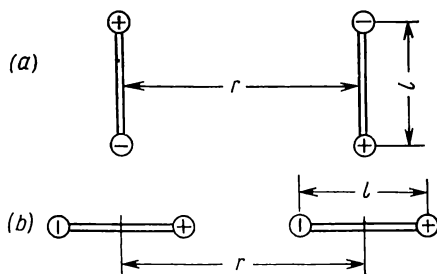


Fig. 98

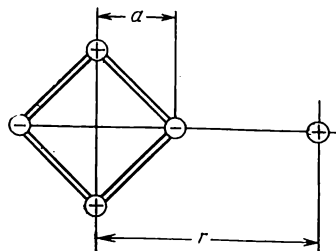


Fig. 99

field, while in a nonuniform field it will be drawn toward the region of greater field intensity (Fig. 97). If the nonuniform field is the field of a point charge, the dipole will be drawn toward this charge. This force of attraction is

$$F = q_0 \left( \frac{q}{r^2} - \frac{q}{(r+l)^2} \right).$$

Assuming the distance between the dipole charges to be small, and reducing the expression to a common denominator, we obtain the following interesting formula if we neglect the quantity  $l^2$  relative to  $rl$  and the quantity  $rl$  relative to  $r^2$ :

$$F = q_0 \frac{2p}{r^3}.$$

Note that the force of interaction between the charge and the dipole decreases more rapidly with increasing distance than the Coulomb force, i.e., it is inversely proportional to the distance cubed.

*Example.* The distance between the H atom and the Cl atom in an HCl molecule is equal to  $1.28 \text{ \AA}$ , and the dipole moment of the molecule is  $p = 6 \times 10^{-18}$  CGS units. Therefore, an electron located at a distance  $r = 10 \text{ \AA}$  from the molecule is attracted to it with a force of  $\sim 6 \times 10^{-6}$  dyne.

**Dipole-Dipole.** Here, it is convenient to solve the problem for the two cases in which the dipoles are arranged as shown in Fig. 98. The exact interaction for-



mulas have the form:

$$F = \frac{2q^2}{r^2} - \frac{2q^2}{r^2 + l^2} \frac{r}{\sqrt{r^2 + l^2}} \quad \text{for arrangement (a)}$$

$$F = \frac{2p^2}{r^2} \times \frac{3r^2 - l^2}{(r^2 - l^2)^2} \quad \text{for arrangement (b).}$$

If the distance between the charges of a dipole is small the above formulas may be replaced by the following approximate expressions:

$$F = \frac{3p^2}{r^4} \quad \text{for arrangement (a)}$$

$$F = \frac{6p^2}{r^4} \quad \text{for arrangement (b).}$$

Thus, the interaction forces are inversely proportional to the fourth power of the distance.

*Example.* Two HCl molecules that are 10 Å apart are attracted with a force  $F \sim 10^{-6}$  dyne in the case of arrangement (a) and with a force  $F \sim 2 \times 10^{-6}$  dyne in the case of arrangement (b).

**Charge-Quadrupole.** Assume that the orientation is as shown in Fig. 99. The interaction force may then be written in the form

$$F = 2q_0q \left( \frac{r^2 + a^2}{(r^2 - a^2)^2} - \frac{r}{(r^2 + a^2)^{3/2}} \right),$$

and the approximate formula for a small quadrupole is  $F = \frac{9q_0qa^2}{r^4}$ . Thus, the force is inversely proportional to the fourth power of the distance.

#### Sec. 94. POLARISATION OF AN ISOTROPIC DIELECTRIC

As we know, when a region containing an electric field created by a system of charges is filled with a homogeneous dielectric, the field intensity and the magnitude of the electric potential are decreased to  $\frac{1}{\epsilon}$  of their original values. On the other hand, the electric displacement and induction remain unchanged, and the capacitance of a condenser increases  $\epsilon$  times its original value. The latter effect is often utilised in the measurement of dielectric constants. Thus, the ratio of the capacitance of a condenser containing a dielectric between its plates to the capacitance of the same condenser without dielectric may be used as a definition of dielectric constant.

Let us now go a step further and inquire into the reasons why the dielectric affects the electric field. The following experiment suggests the explanation for this phenomenon.

Consider a parallel-plate condenser connected to a source of voltage. The electric charge density on the condenser plates and, hence, the number of  $D$ -lines per unit area are uniquely determined by the electric field intensity, i.e.,  $\sigma = \frac{E}{4\pi}$ . Let us now fill this condenser with a homogeneous dielectric. The relation between the electric field intensity and the charge density on the condenser plates is then expressed by the equation  $\sigma = \frac{\epsilon E}{4\pi}$ , i.e., the flux density (or  $D$ -lines) increases. In this experiment, the electric field intensity cannot change, for it is equal to the potential difference divided by the distance between the plates. Therefore,

the charge density on the condenser plates changes, i.e., it increases  $\epsilon$  times its original value. This increase may be observed experimentally. Thus, as we fill the condenser with the dielectric, the voltage source adds charge to the condenser. By measuring the electric current and the time of flow, one can show that the quantity of electricity added per unit condenser area is

$$\frac{\epsilon E}{4\pi} - \frac{e}{4\pi} = \frac{\epsilon - 1}{4\pi} E.$$

As we remove this dielectric, the additional charge returns to the source and the additional force lines disappear. To explain the additional attraction of charge to the condenser plates, one must assume that charges of opposite sign, having a density  $\sigma = \frac{\epsilon - 1}{4\pi} E$ , are formed on the dielectric surface next to the condenser plates.

The surface charge of the dielectric may be explained if we assume that the dielectric consists of bound pairs of positive and negative charges that cannot move through the body, but can move relative to each other, forming thereby a dipole moment in each unit volume of the dielectric. This transformation of an electrically neutral system of charges into a system having a dipole moment is called *polarisation*, and the dipole moment vector of a unit volume of dielectric is called the *polarisation vector*  $\mathbf{P}$ .

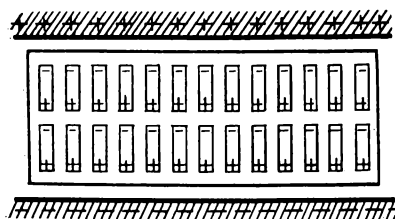


Fig. 100

Polarisation of a dielectric does not produce volume charge. The numbers of positive and negative charges per unit volume remain equal to each other after displacement. However, polarisation does produce a charge on the dielectric surface. This is illustrated schematically in Fig. 100. The density of this charge is equal to the value determined above, i.e.:

$$\sigma_{POL} = \frac{\epsilon - 1}{4\pi} E.$$

In our discussion, we have considered a dielectric adjacent to the plate of a parallel-plate condenser. However, the situation is the same for a conductor surface of any shape. Moreover, it turns out that the above expression for  $\sigma_{POL}$  is of general validity for surfaces perpendicular to lines of force. Thus, in every case,

$$\sigma = \frac{\epsilon - 1}{4\pi} E_n,$$

where  $E_n$  is the projection of the intensity on the normal to the surface. This formula is applicable to any real or imaginary boundary in a dielectric.

A polarised charge (also often called a bound charge) may be expressed in terms of the dipole moment of a unit volume. When a field is applied in the case of isotropic bodies, the displacement of the bound charges occurs along the electric lines of force. Therefore, the polarisation vector is parallel to the intensity vector. From a dielectric plate, let us cut out a cylindrical rod having a base  $S$  and a length  $l$ . Owing to polarisation, equal and opposite bound charges will accumulate at the ends of the cylinder. The dipole moment of the rod is equal, by definition, to the product of the charge  $\sigma S$  and the distance  $l$  separating the charges of the dipole, i.e.,  $p = \sigma_{POL} S l$ . The dipole moment of a unit volume is  $|\mathbf{P}| = \sigma_{POL}$ . It has been assumed in this calculation that the base of the cylinder is perpendic-

cular to the polarisation direction. If the base is inclined at an angle  $\varphi$  with respect to this position, the charge density on the ends of the rod will decrease as the cosine of the inclination angle. Thus, in the general case, the following relationship holds:

$$\sigma_{POL} = P_n, \quad \text{where} \quad P_n = P \cos \varphi.$$

We are now able to establish a relationship between the polarisation vector and the electric field intensity vector. Combining the last formula with the expression for the density of bound charges discussed at the beginning of this article, we obtain for any direction  $n$ :

$$P_n = \frac{\epsilon - 1}{4\pi} E_n.$$

Thus, if the dielectric constant does not depend on the intensity, a linear dependency exists between the vectors  $\mathbf{P}$  and  $\mathbf{E}$ :

$$\mathbf{P} = \alpha \mathbf{E}.$$

The expression  $\alpha = \frac{\epsilon - 1}{4\pi}$  is usually called *the polarisability*. For water  $\alpha = 6.38$  and for glass  $\alpha = 0.48$ .

Since  $\mathbf{D} = \epsilon \mathbf{E}$ , the relationship between the vectors  $\mathbf{D}$ ,  $\mathbf{E}$  and  $\mathbf{P}$  may be expressed in the form

$$\mathbf{D} = \mathbf{E} + 4\pi \mathbf{P}.$$

When the medium is homogeneous, the vectors  $\mathbf{D}$ ,  $\mathbf{E}$  and  $\mathbf{P}$  are parallel.

#### Sec. 95. POLARISATION OF CRYSTAL SUBSTANCES

Hitherto we have considered the behaviour of a substance that is characteristic of an amorphous or finely crystalline body, or of a monocrystal in certain special orientations relative to the field. However, if a plate is cut from a monocrystal at an arbitrary angle to the crystal faces and if the plate is then placed between the conductors of a condenser, the following effect may be observed: The plate becomes polarised perpendicular as well as parallel to the lines of force, so that  $\mathbf{P}$  is not parallel to  $\mathbf{E}$ . Therefore, in this case,  $\mathbf{D}$  is also not parallel to the field intensity.

In monocrystals, the inclination direction of a free electric charge ( $\mathbf{E}$ ) does not coincide with the direction of the normal to the surface oriented in such a manner that a maximum charge ( $\mathbf{D}$ ) is induced on it. The relationship between  $\mathbf{D}$  and  $\mathbf{E}$  becomes more complex and in order to find  $\mathbf{D}$  in terms of  $\mathbf{E}$ , or vice versa, it is insufficient to merely know the dielectric constant. In any monocrystal, three directions (main axes) in which  $\mathbf{D} \parallel \mathbf{E}$  may be found. If we know  $\epsilon$  for these three directions, the relation between  $\mathbf{D}$  and  $\mathbf{E}$  for any arbitrary orientation of the crystal in a field may then be established. How are the vectors  $\mathbf{D}$ ,  $\mathbf{E}$  and  $\mathbf{P}$  related in this case? It turns out that the equation  $\mathbf{D} = \mathbf{E} + 4\pi \mathbf{P}$ , introduced in the previous article for the case of parallel vectors, is also valid when the vectors cease to be parallel. There is another difference between crystals and amorphous bodies in regard to their dielectric properties, namely, a relatively small class of bodies belonging to crystalline substances possess hysteretic properties. Since these properties were first discovered in Seignette (Rochelle) salt, such substances are called seignette-electric materials. Their characteristics will now be discussed (see also p. 468).

Place a seignette-electric material (for simplicity assume that we are dealing with a powder or crystal oriented in the field so that  $\mathbf{D} \parallel \mathbf{E}$ ) between the conductors of a condenser. Let us vary the voltage between the condenser conductors, and hence the field intensity  $E = \frac{U}{d}$ , and measure the charge density  $\sigma$  on the condenser conductors, which for  $\mathbf{D} \parallel \mathbf{E}$  yields the magnitude of the electric induction  $D$ . The magnitude of the induction  $D$  increases as  $E$  increases, but it is not directly proportional to  $E$ , for  $D$  begins to increase less until, finally, saturation sets in. Clearly, saturation of  $D$  corresponds to saturation of polarisation. Let us now begin

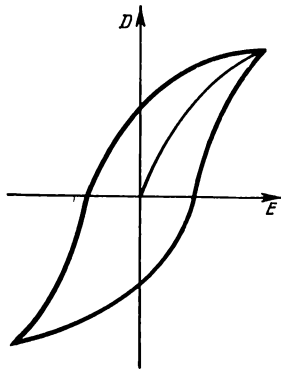


Fig. 101

to decrease the emf between the conductors. Displacement and polarisation begin to decrease and the curve follows a downward path, but not the same one taken during the period of rise. As a result, when the emf is completely removed ( $E = 0$ ), the induction and polarisation in the dielectric are not equal to zero. The dielectric becomes similar to a permanent magnet. It will have "north" and "south" electric poles and will behave like a large permanent dipole.

The subsequent behaviour of seignette-electric materials is evident from the hysteresis loop shown in Fig. 101. To "de-electrify" the dielectric, the polarity of the emf on the condenser conductors must first be reversed. Then, by increasing  $E$  in this new direction, we can depolarise the dielectric. A further increase in  $E$  again electrifies it, but with opposite polarity. Finally,

saturation sets in again and the process may be repeated in the reverse direction.

Why is this effect called hysteresis? The word is derived from the Greek and means "to lag". The loop illustrated in the figure shows that the values of  $D$ , as well as of  $P$  and  $\epsilon$ , depend on the past state of the sample, i.e., on its history.

Every crystal that does not possess a centre of symmetry in a number of its elements of symmetry (see p. 468) possesses an interesting property, namely, its dimensions change upon application of an electric field. This phenomenon is known as *electrostriction*.

Thermodynamic considerations show that if an electric field produces a deformation, the deformation in turn will produce polarisation. This is known as the *piezoelectric effect*. Applications of the piezoelectric effect were briefly discussed in Part I.

## Sec. 96. FINITE DIELECTRIC BODIES IN AN ELECTRIC FIELD

The following questions may arise regarding a finite nonconducting body located in an electric field. What forces and moments of force act on such a body? How is the field distorted by the presence of the dielectric?

A dielectric body placed in a field becomes polarised and acquires a certain dipole moment. Therefore, the behaviour of such a body in an electric field does not, generally speaking, differ from the behaviour of a dipole. If the polarisation vector is directed at an angle to the field intensity, the orientation of the dielectric will be unstable. A moment of force will act on the body, which will tend to turn the body until the vectors  $\mathbf{P}$  and  $\mathbf{E}$  are parallel.

Thus, a dielectric body placed in a given uniform electric field assumes a definite equilibrium orientation that depends on the form of the body. Let us illustrate this for the case of a dielectric bar.

Experiments show that the equilibrium orientation is the one for which the longitudinal axis is parallel to the lines of force. Why is this so? Is it not true that the bar does not have fixed poles? Fig. 102 illustrates the reason for this peculiar behaviour. The forces acting on the bound charges of the depicted rectangular bar may be reduced to four forces acting on four surfaces of the bar. We see that the forces acting on the longitudinal surfaces almost balance each other, while the forces acting on the lateral surfaces form a couple of forces that orients the bar parallel to the lines of force.

If the body is in a nonuniform field, then, in addition to the moment of force, there will exist forces tending to pull the dielectric toward the region of greater field intensity. This phenomenon may be vividly demonstrated by making a dielectric fluid rise in a tube upon applying voltage to a condenser. The forces making bits of paper cling to a glass or ebonite rod rubbed with fur or leather are of the same kind as those acting on a dipole in a nonuniform field.

Let us now turn to the question relating to the distortion of an electric field due to the presence of a dielectric body. First, we shall show that the general laws for an electric field lead to an important relation between the values of the electric field on either side of the boundary between dielectrics.

The electric field intensity vectors at two neighbouring points located on opposite sides of the boundary between dielectrics having permittivities  $\epsilon_1$  and  $\epsilon_2$  differ in magnitude as well as in direction. Let us resolve these vectors into components parallel and normal to the boundary. It can be asserted that the field parallel to the boundary is the same on both sides. If this were not the case, i.e., if the field on one side were greater than the field on the other side, it would be possible to create a perpetual engine by moving charges along the boundary where the field is less and then allowing the charge to move on the other side of the boundary (where the field is greater) under the action of the electric field forces. Therefore, the tangential components of the intensity on both sides of the boundary must be equal:

$$E_{t_1} = E_{t_2}.$$

We shall use the Gauss-Ostrogradsky law to determine the normal components of the intensity at the boundary between (or interface of) two media. Construct an auxiliary surface in the form of an infinitely thin disk so that the parallel surfaces of the disk lie on opposite sides of the boundary. Since there is no charge inside such a disk, the net outward flux through the disk is equal to zero and, therefore, the flux through each end is the same. This requires that the normal components of the induction vectors be equal to each other, i.e.,  $D_{n_1} = D_{n_2}$ . Hence, in terms of field intensities, we obtain

$$\epsilon_1 E_{n_1} = \epsilon_2 E_{n_2}.$$

Thus, the normal components of the intensity vectors are inversely proportional to their permittivities.

Fig. 103 shows that in passing from a medium of lower permittivity to one of higher permittivity the flux lines are deflected away from the normal to the boundary. This means that the number of flux lines passing through a unit area increases.

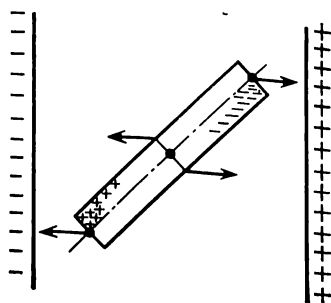


Fig. 102

We are unable to determine the distortion produced in an electric field when a dielectric having a particular shape is introduced into this field. This problem is difficult even when the field is uniform to begin with. If a body of arbitrary shape is placed in such a field, the field becomes nonuniform not only near the body, but inside the body as well.

Interesting exceptions are ellipsoids, a broad class of bodies including spheres, flattened ellipsoids that practically do not differ from plates, and extended ellipsoids that are akin to cylindrical bodies. In mathematical physics it is shown that the field inside an ellipsoid is uniform as indicated in Fig. 104. Applying the law

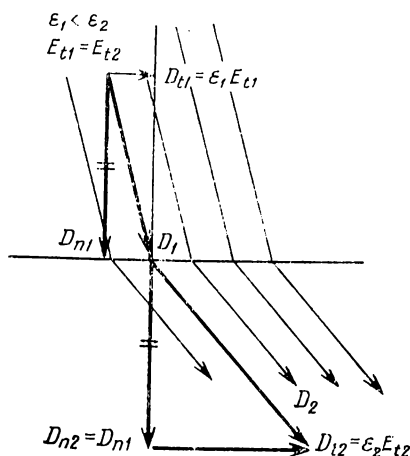


Fig. 103

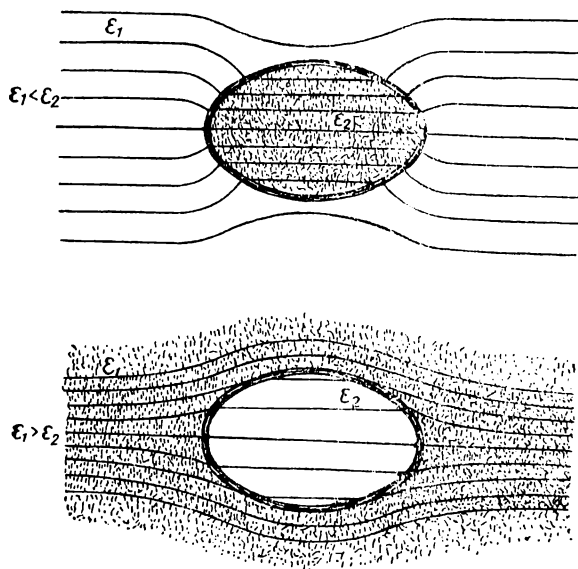


Fig. 104

of flux refraction, we obtain the typical fields illustrated for a denser body in a less dense medium ( $\epsilon_1 < \epsilon_2$ ) as well as for the reverse case ( $\epsilon_1 > \epsilon_2$ ). Examples are a glass ellipsoid in air and an air bubble in glass, respectively.

It can be shown that if a symmetrical dielectric body is immersed in a uniform field  $E_0$  in vacuum, then  $E_i$  is related to the field inside the dielectric as follows:

$$E_i = E_0 - NP,$$

where  $P$  is the polarisation vector and  $N$  is a coefficient depending only on the shape of the body. In the case of magnetic phenomena, it is customary to call the latter the demagnetisation coefficient (see p. 221).

Since in most cases  $P = \frac{\epsilon - 1}{4\pi} E_i$ , we obtain the following expression after simple conversion:

$$E_i = \frac{\epsilon_0}{1 + (\epsilon - 1) \frac{N}{4\pi}}.$$

The dielectric constant is always greater than unity. Hence, the field intensity inside the dielectric is always less than the field intensity present at this location before the dielectric was introduced into the field.

The coefficient  $N$  for a flat plate perpendicular to the field is equal to  $4\pi$ . This is the maximum value for  $N$  and the resulting decrease in field to  $\frac{1}{\epsilon}$  of its original value agrees with the result obtained earlier for a homogeneous medium. Let us take another extreme case—a cylinder whose axis is oriented parallel to the field. Here,  $N = 0$ , i.e., the field does not decrease in such a body. In all other cases, the decrease in the field intensity depends on the dielectric constant. For a sphere,  $N = \frac{4\pi}{3}$  and, therefore,  $E_i = \frac{3E_0}{\epsilon + 2}$ . For a cylinder whose axis is oriented perpendicular to the field,  $N = \frac{4\pi}{2}$ .

The field intensity  $E$  decreases because the bound charges create a field of opposite direction.

As regards the induction vector field, the bound charges affect it only indirectly. Thus, the number of  $D$ -lines remains unchanged when a dielectric is immersed in the field. However, due to flux refraction, the induction inside the dielectric increases.

# Magnetic Fields

## Sec. 97. MAGNETIC MOMENT

Magnetic fields act on currents, moving charged bodies or particles and magnetised bodies. A variety of instrument exist for determining the properties of a magnetic field. The most convenient way of characterising the properties of such a field is to describe its mechanical action on a current circuit. It is quite feasible to construct a wire circuit of very small area, which enables us to measure the magnetic field quite accurately. Thus, a "test" current circuit plays the same role in magnetic field theory as a "test" charge does in electric field theory.

Experiments with such a device lead us to the following basic conclusions. At each point of the field, a circuit that is free to rotate assumes a definite equilibrium position. The position of stable equilibrium is described not only by the orientation of the circuit axis in space, but by the orientation of a definite side of the circuit, e.g., the side for which the current will appear to be flowing counter-clockwise as viewed by an observer on that side of the circuit. Let us call this side

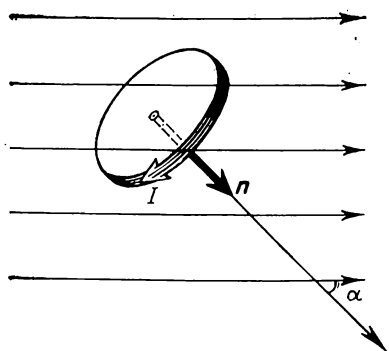


Fig. 105

positive or north and agree to draw the normal to the circuit so as to form a right-hand screw-system with the current direction. Thus, the normal emerges on the positive (or north) side of the circuit.

If the behaviour of current circuits is compared with that of magnetic needles, one observes that the normal to a circuit in stable equilibrium points in the same direction as a magnetic needle. Thus, the basic definition is not contradicted if we call the direction of the normal to a free test circuit the direction of the magnetic field.

A torque will act on a test circuit that deviates from the equilibrium position (Fig. 105). Moreover, the deviation of the circuit from equilibrium is uniquely described by the deviation of the normal to the circuit from the direction of the field. It turns out that the sine of angle  $\alpha$  and the torque  $N$  are proportional to each other, i.e.,  $N \sim \sin \alpha$ . Furthermore, for a particular angle  $\alpha$ , the torque is proportional to the product of the circuit area  $S$  and the current  $I$  flowing in the circuit. Decreasing the area by a certain factor results in the same change in torque as a decrease in current by the same factor.

It follows from the above that the magnetic behaviour of a circuit depends on the orientation of the normal to the circuit and on the magnitude of the product  $IS$ . This can be expressed by means of a single vector quantity—the so-called *magnetic moment of the ring current*. In electrical engineering, where SI units are used, the magnetic moment is generally designated by the vector  $\mathbf{M} = IS\mathbf{n}$ , where  $\mathbf{n}$  is the unit normal vector. In the Gaussian system of units used in physics, a constant of proportionality  $\frac{1}{c}$  enters into this formula, i.e.,  $\mathbf{M} = \frac{1}{c} IS\mathbf{n}$ , where  $c$  is the velocity of propagation of electromagnetic waves in vacuum. The introduction of



a numerical coefficient, which is moreover dimensional, may appear to be an unnecessary complication. However, other formulas are thereby simplified. This will become clear to the reader later on.

Experiments with a test circuit show that  $N = BM \sin \alpha$ , where  $B$  is a constant of proportionality. The volume of  $B$  varies from field to field and for different points in space of a particular field. This formula shows that  $B$  is equal to the maximum torque acting on a unit test circuit ( $M = 1$ ). We call this coefficient  $B$ , which characterises the magnetic field, *the magnetic induction*. The vector quantity whose direction is that of the magnetic field and which is numerically equal to  $B$  is known as *the magnetic induction vector*.

If the torque is described by a vector directed along the axis of rotation (in accordance with the right-hand screw rule), the formula for the torque may be written in vector form as follows  $\mathbf{N} = [\mathbf{MB}]$ .

When  $N = 0$ ,  $\mathbf{M}$  is parallel to  $\mathbf{B}$ . This means that a current circuit tends to become so oriented in a magnetic field that the directions of the magnetic moment and the field coincide. The magnetic moment acting on a body is a maximum when it is perpendicular to the direction of the field. For a circuit, this means that the plane of the loop of wire is parallel to the flux lines.

Having determined the magnetic field by means of a current circuit whose magnetic moment has been calculated from measurements of the current and the area, we can then do the reverse, namely, use the formula  $\mathbf{N} = [\mathbf{MB}]$  to determine the magnetic moment of systems whose currents cannot be measured. Moreover, we can transfer the concept of magnetic moment to systems in which the concept of a circular current has no meaning. This is precisely the case when the physicist refers to the magnetic moment of an electron or nuclear particle. The magnetic moment of a magnetic needle is also a concept that cannot be broken down. However, after having discussed certain special effects of the medium, we shall return once again to the magnetic moment of a permanent magnet. In any case, the magnetic moment of a system located in vacuum can always be determined by the above formula for the torque.

A body possessing a magnetic moment requires the expenditure of work to turn it from its equilibrium position. In the case of a body turned through a small angle  $\alpha$ , the work of rotation may be expressed in the form

$$N d\alpha = BM \sin \alpha d\alpha = -d(BM \cos \alpha).$$

The deviation of a body from the equilibrium position is associated with the accumulation of a "potential" energy  $U = -BM \cos \alpha$ . This product is the scalar product of two vectors. Hence,  $U = -\mathbf{BM}$ .

In the equilibrium position, the potential energy is a minimum and equal to  $-BM$ . When the magnetic moment is turned through an angle of  $90^\circ$  the potential energy increases to zero. Finally, when the magnetic moment is directed oppositely to the field (position of unstable equilibrium), the potential energy is a maximum and equal to  $+BM$ .

*Examples.* 1. The magnetic moment of the nucleus of a hydrogen atom (nuclear magneton) is  $0.505 \times 10^{-23}$  CGS unit. The magnetic moment of an electron (Bohr magneton) is  $0.927 \times 10^{-20}$  CGS unit  $= 9.27 \times 10^{-24}$  A  $\times$  m<sup>2</sup>.

2. An electric current of 1 A flowing in a loop whose area is 50 cm<sup>2</sup> creates a magnetic moment of  $5 \times 10^{-3}$  A  $\times$  m<sup>2</sup>  $= 5$  CGS units.

3. In the CGS system of units magnetic induction is measured in gauss (G) while in the SI system  $B$  is measured in teslas (T) and has the dimension of V  $\times$  sec/m<sup>2</sup>; 1 T =  $10^4$  G. In the case of the magnetic field of the Earth,  $B = 0.49$  G.

4. The magnetic induction in the air gap of a powerful electric generator attains a value of several thousand gauss. Academician P. L. Kapitsa has obtained pulsed magnetic fields in which  $B \sim 10^5$  G = 10 T.

## Sec. 98. AMPÈRE FORCE

The torque acting on a current-carrying circuit is clearly the resultant of the forces exerted on every part of the conductor in which current flows. We can experimentally establish the relation for the force acting on a current element. For this purpose, it is necessary to isolate a part of the circuit—e.g., by means of mercury contacts. This part is then able to move under the action of a force. Utilising the tension of a spring to counterbalance the displacement, one can measure the magnetic force (Fig. 106).

Ampère first established the relation for the force acting on a current element of small length. This relation has the following form:

$$d\mathbf{F} = \frac{I}{c} [d\mathbf{l}, \mathbf{B}] \quad \text{i.e.,} \quad dF = \frac{I}{c} dl \times B \sin \angle d\mathbf{l}, \mathbf{B}.$$

The vector notation here is suggestive of the familiar left-hand rule. The force acting on an element of wire length is always perpendicular to the plane passing

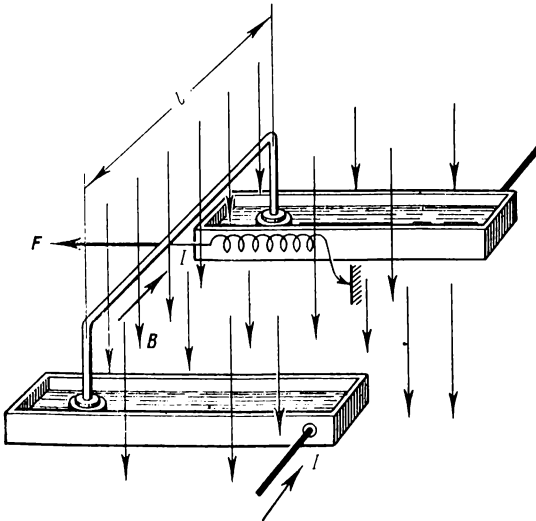


Fig. 106

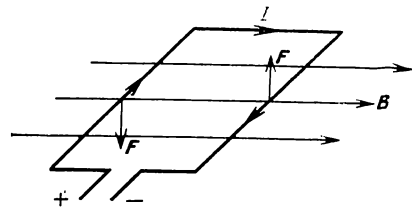


Fig. 107

through the current and the magnetic induction vector at this location. To determine the sense of the force, note from which side the rotation of vector  $d\mathbf{l}$  toward vector  $\mathbf{B}$ , through the smallest angle, appears counterclockwise. This is the positive side in a right-hand screw system and the force vector then points toward the observer. The force has a maximum value when the current element is perpendicular to the vector field. When the wire element is parallel to the flux lines, the force is equal to zero.

The above formulas are in the form used in physics, i.e., valid for the CGS system of units. In the form used in the SI system, the coefficient  $\frac{1}{c}$  is absent and the formula for the Ampère force is

$$d\mathbf{F} = I [d\mathbf{l}, \mathbf{B}].$$

To determine the magnitude of the force acting on a piece of wire of finite length, one must integrate the above expression for the force:

$$F = \frac{1}{c} \int [d\mathbf{l}, \mathbf{B}].$$

In the simple case of a rectilinear piece of wire of length  $l$ , located in a uniform magnetic field  $B$ , Ampère's law may be directly applied in the form

$$F = \frac{1}{c} IlB \sin \angle \mathbf{l}, \mathbf{B}.$$

A perfectly natural relationship exists between Ampère's law and the torque expression derived in the preceding article. We shall consider only the simple case of a rectangular loop oriented parallel to the flux lines in a uniform magnetic field (Fig. 107). Two sides of the loop are perpendicular and the other two sides are parallel to the flux lines. Therefore, the forces acting on the wire elements may be reduced to the two shown in Fig. 107. These forces are equal and in accordance with Ampère's law may be written in the form  $F = IlB$ . As can be seen from the figure, the Ampère forces create a torque  $N = IlBd$ . But since  $ld = S$  is the area of the loop, we obtain  $N = ISB = MB$ , which is the same as the formula for the torque derived in the preceding article. We leave it to the reader to derive a more general proof.

*Example.* The force acting on a conductor whose length is 3 m and through which a current of 50 A flows in a field of 3,000 G = 0.3 T is  $F = Bil = 0.3 \times 50 \times 3 = 45$  N. In the case of a rotor diameter of  $\sim 1$  m, the torque acting on the loop is  $\sim 45$  N  $\times$  m. These values are of the order of magnitude of the parameters of a large electric motor. In an electrical measuring instrument, a force  $F = 2 \times 10^{-6}$  N = 2 dynes acts on a conductor of length 2 cm through which a current of 0.01 A flows in a field of 100 G. For a loop diameter of  $\sim 1$  cm, a torque of  $\sim 2 \times 10^{-7}$  N  $\times$  m acts on the loop.

#### Sec. 99. FORCE ACTING ON A MOVING CHARGE

We may go a step further and consider the magnetic forces acting on currents as forces applied to elementary particles of electricity.

Electric current is simply a flow of electric charges. If  $e$  is the particle charge,  $v$  the particle velocity and  $n$  the particle concentration (i.e., the number per

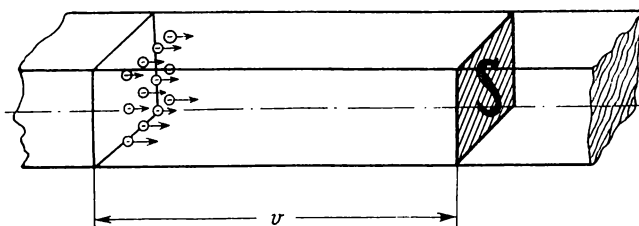


Fig. 108

unit volume), then the expression for the current intensity may be expressed in the form  $I = nevS$ . Thus, all the particles in a volume  $vS$  pass in 1 sec through the wire cross-section  $S$ , i.e., the quantity of electricity flowing is equal to  $nevS$  (Fig. 108). Substituting this expression in the formula for Ampère's law, we obtain:

$$d\mathbf{F} = \frac{e}{c} [v\mathbf{B}] nS d\mathbf{l}.$$

But  $nS \, dl$  is the number of particles in the conductor volume under consideration. Thus, the force acting on one particle is

$$f = \frac{e}{c} [vB].$$

This force is sometimes called *the Lorentz force*—in honour of Lorentz, the distinguished physicist who contributed so much toward the development of the theory of electrons.

The above expression for the force (we shall restrict ourselves to the form used in physics, i.e., with the coefficient  $\frac{1}{c}$ ) immediately provides the answer to a number of very interesting questions regarding the nature of motion of electric particles (e.g., electrons and protons) in a magnetic field. The force acting on a moving particle is perpendicular to the flux lines and to the particle velocity vector. If a particle moves parallel to the flux lines, no force is exerted on it. On the other hand, the force is a maximum when the motion occurs in a plane that is perpendicular to the flux lines. In this case, we obtain  $f = \frac{1}{c} evB$ .

If the field is uniform, an electric particle moving perpendicular to the field will describe a circle, for according to the fundamental law of mechanics this is the nature of the motion under the action of a constant force directed at right angles to the motion. We shall return to the problem of particle motion in a magnetic field later.

*Example.* Electrons in a cathode-ray tube accelerated by a potential difference of 70 V acquire a velocity of  $5 \times 10^8$  cm/s. Upon entering a magnetic field of 500 G at right angles to the field, each electron experiences a Lorentz deflecting force of  $f = \frac{1}{c} evB = 4 \times 10^{-11}$  dyne. Under the action of this force, the electron begins to move in a circular orbit of radius  $R$ , where  $R$  is determined from the relation  $f = \frac{mv^2}{R}$ . Hence,  $R = 5.6$  cm.

## Sec. 100. MAGNETIC FIELDS CREATED BY PERMANENT MAGNETS

Every permanent magnet has two poles.\* The flux lines are directed outwardly at the north pole and inwardly at the south pole. Imagine a surface constructed so that it encloses the north pole. We can then determine the total number of lines passing outwardly through this surface. By analogy with the corresponding electric quantity, this number is called *the magnetic flux* and is designated by the letter  $\Phi$ . The flux through an elemental area perpendicular to the flux lines is equal to  $d\Phi = BdS_{\perp}$ . Thus, through any arbitrary area,  $d\Phi = BdS \cos \alpha$ , where  $\alpha$  is the angle formed with the flux lines by the normal to the area; and through the surface  $S$ ,  $\Phi = \int B \cos \alpha \, dS$ . Finally, through the closed surface,  $\Phi = \oint B \cos \alpha \, dS$ .

The flux  $\Phi_N$ , directed outwardly at the north pole and inwardly at the south pole, is the fundamental characteristic of a magnet. The stronger the magnet, the greater  $\Phi_N$ . This somewhat justifies the designation "quantity of magnetism"—which is only of historical significance—for the quantity proportional to the flux, namely,  $m = \frac{1}{4\pi} \Phi$ . Sometimes  $m$  is called the magnetic mass, but this term is even less appropriate. In electrical engineering units,  $m = \Phi$ .

\* The creation of magnets with any number of *pairs* of poles is also conceivable.

If the poles of a magnet are small (e.g., in the case of a magnetic needle), the flux lines close to the poles are directed radially.

Using the Gauss-Ostrogradsky theorem,

$$\oint D \cos \alpha dS = 4 \pi q,$$

we derived the formula for the electric induction of an isolated charge, namely,  $D = \frac{q}{r^2}$ . Clearly, an "isolated" magnetic pole should yield a magnetic induction satisfying the analogous equation:

$$B = \frac{m}{r^2}, \quad \text{since} \quad \oint B \cos \alpha dS = 4\pi m \text{ (CGS),}$$

or

$$B = \frac{m}{4\pi r^2}, \quad \text{since} \quad \oint B \cos \alpha dS = m \text{ (SI).}$$

To be sure, "isolated" magnetic poles do not exist. The above formula has meaning only for long magnets having point poles, and then only close to the poles. Never-

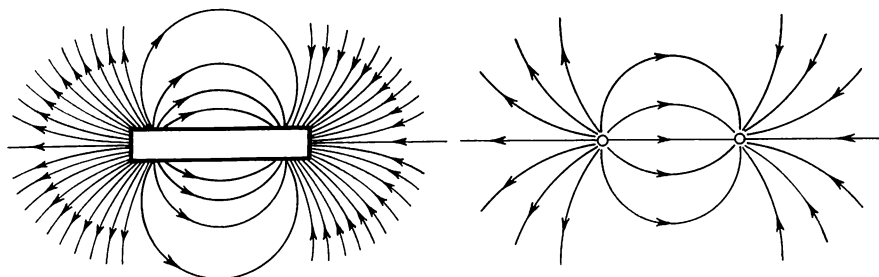


Fig. 109

theless, this method of dealing with the pole of a permanent magnet is fully justified in practice. This can be clearly demonstrated by means of the expression for the field of a bar magnet considered as a magnetic dipole whose two poles  $m$  are separated by a distance  $l$ . Fig. 109 shows the field of a bar magnet and the ideal field based on the formula

$$B = \frac{m}{r_1^2} \frac{r_1}{r_1} - \frac{m}{r_2^2} \frac{r_2}{r_2},$$

where  $r_1$  and  $r_2$  are the distances from the poles to the point under consideration. The fields are seen to be very similar.

Calculations yield good results for the field at large distances from the magnet. Thus, if the distances  $r_1$  and  $r_2$  are large relative to the magnet length  $l$ , the distance between the poles of the magnetic dipole, we are fully justified in considering the poles as points. The calculations are exactly the same as the corresponding calculations for electric interactions. Let us compare, for example, the values of the magnetic induction created by a bar magnet at a distant point along the magnet axis and at a distant point perpendicular to the axis. In the first case, we obtain

$$B = \frac{m}{r^2} - \frac{m}{(r+l)^2} = \frac{2ml}{r^3} = \frac{2M}{r^3},$$

where  $M = ml$  is called the magnetic moment of the permanent magnet. In the second case (Fig. 110).

$$B = 2 \frac{m}{r^2} \cos \omega = \frac{M}{r^3}.$$

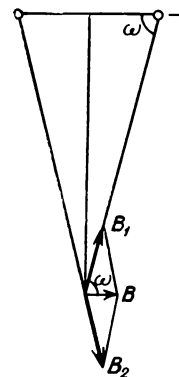


Fig. 110

Thus, the field along the axis is twice as large.

In the SI system of units, the last two formulas assume, respectively, the following form:

$$B = \frac{M}{2\pi r^3} \quad \text{and} \quad B = \frac{M}{4\pi r^3}.$$

*Example.* Let us calculate the magnetic induction created by a bar magnet of length  $l = 10$  cm at a distance  $r = 1$  metre from the magnet, measured along the axis. The magnet cross-section is  $S = 3$  cm<sup>2</sup> and the induction in the magnet is 500 G.

The magnetic flux in the magnet, which is the same as the outward flux from the pole, is  $\Phi = 500 \times 3 = 1,500$  maxwells (Mx). Thus, a "magnetic mass"  $m = \frac{1,500}{4\pi} = 120$  CGS units is concentrated at the magnet pole. The magnetic moment of the magnet is

$$M = ml = 120 \times 10 = 1,200 \text{ ergs/G (CGS units)}.$$

Then, for the magnetic induction, we obtain

$$B = \frac{2M}{r^3} = \frac{2 \times 1,200}{(100)^3} = 2.4 \times 10^{-3} \text{ G}.$$

#### Sec. 101. MAGNETIC FIELD INTENSITY

Let us consider the interaction of an isolated magnetic pole and a current element (Fig. 111). The magnetic pole creates a field  $\mathbf{B}$  at the location of the electric current. Therefore, in accordance with Ampere's law, the force acting on the current element is

$$d\mathbf{F} = \frac{1}{c} I [d\mathbf{l}, \mathbf{B}].$$

In place of the magnetic induction, we can substitute the expression for a point pole. Since the field is directed along the radius, we obtain the following expression for the interaction force:

$$d\mathbf{F} = \frac{m}{cr^2} I \left[ d\mathbf{l}, \frac{\mathbf{r}}{r} \right]$$

or

$$dF = \frac{mI}{cr^2} dl \times \sin \angle d\mathbf{l}, \mathbf{r}.$$

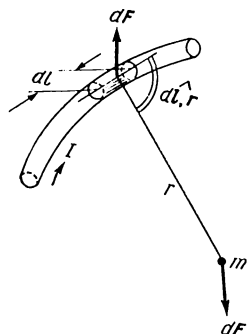


Fig. 111

It is quite natural to assume that the force with which a current element acts on a magnetic pole is represented by the same formula, except that the direction of the force is reversed. This assumption cannot be directly verified experimentally since an isolated pole and an isolated element of constant current do not exist. However, we can verify the validity of the above statement by integrating the interaction forces in actual cases. It turns out that theory and experiment agree.

Thus, the force exerted by a current element on a magnetic pole may be written in the form

$$d\mathbf{F} = \frac{m}{cr^2} I \left[ d\mathbf{l}, \frac{\mathbf{r}}{r} \right]$$

or, in the SI, i.e., without the coefficient  $\frac{1}{c}$  and replacing  $m$  by  $\frac{m}{4\pi}$ :

$$d\mathbf{F} = \frac{m}{4\pi r^2} I \left[ d\mathbf{l}, \frac{\mathbf{r}}{r} \right].$$

A minus sign does not appear in this formula because a reversed radius vector has been assumed. The direction of  $\mathbf{r}$  is always taken as the direction from the field source to the point of observation. Therefore, in the case of the force acting on the current,  $\mathbf{r}$  was assumed to be directed from the pole to the current element. Now, when the force is exerted by the current on the pole, the radius vector  $\mathbf{r}$  is assumed to be directed from the current element to the pole.

The force acting on a unit magnetic pole is called *the magnetic field intensity*:

$$d\mathbf{H} = \frac{d\mathbf{F}}{m}.$$

Thus, our discussion has shown that the magnetic field intensity created by a current element is given by the formula:

$$d\mathbf{H} = \frac{I}{cr^2} \left[ d\mathbf{l}, \frac{\mathbf{r}}{r} \right].$$

In the SI system of units, this formula for the magnetic field intensity created by a current has the form

$$d\mathbf{H} = \frac{I}{4\pi r^2} \left[ d\mathbf{l}, \frac{\mathbf{r}}{r} \right].$$

Thus, a magnetic field may be characterised in two different ways, namely, by the induction vector and the intensity vector. The former measures the action of the magnetic field on a current and the latter measures the action of the field on a magnet.

In practice, it is easier to reduce the measurement of intensity to the measurement of the torque acting on a magnetic needle (Fig. 112). Such a needle located in a uniform field is subject to the action of a couple of forces, where the magnitude of the force is equal to  $mH$  and the arm of the couple is equal to  $l \sin \alpha$ . Hence, for the torque, we obtain the expression

$$N = MH \sin \alpha$$

In vectorial form  $N = [MH]$ , where  $\mathbf{M} = m\mathbf{l}$  is the magnetic moment of the needle. It is seen that this formula is very similar to that for the torque acting on a current-carrying circuit.

The relationship existing between the magnetic field intensity and the magnetic induction can be determined experimentally. It turns out that in all cases, except in the case of anisotropic bodies, the intensity and induction vectors are parallel to each other. This means that the magnetic needle and the axis of the test circuit are always parallel. Moreover, in all cases, except in the case of ferromagnetic substances, a simple linear relationship exists between  $\mathbf{H}$  and  $\mathbf{B}$ , namely  $\mathbf{B} = \mu_0 \mu \mathbf{H}$ , where  $\mu_0$  is a universal constant (the so-called *magnetic permeability of vacuum*) and  $\mu$  is a coefficient characterising the medium (*the relative magnetic permeability of the medium*).

In the CGS system,  $\mu_0 = 1$ . This yields the same dimensions for the magnetic induction and the intensity. The price that must be paid for this identity, however, is the introduction of the dimensional coefficient  $\frac{1}{c}$  in the formula for Ampère's law. In the SI system, the magnetic permeability of vacuum is

$$\mu_0 = 4\pi \times 10^{-7} \text{ J}/(\text{A}^2 \times \text{m}).$$

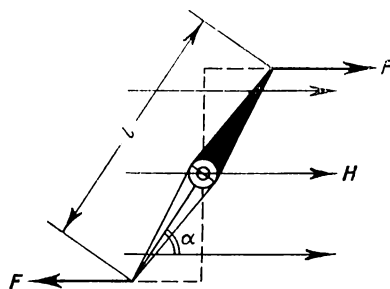


Fig. 112

## Sec. 102. INTERACTIONS OF CURRENTS AND MAGNETS

The relations considered in the preceding sections enable us, in principle, to calculate the interactions of any magnetic system. We have at our disposal formulas for the forces and torques acting on devices by a magnetic field of any origin:

Action on a current		Action on a magnet
CGS	SI	
$\mathbf{F} = \frac{I}{c} [d\mathbf{l}, \mathbf{B}]$	$\mathbf{F} = I [d\mathbf{l}, \mathbf{B}]$	$\mathbf{F} = m\mathbf{H}$
$N = [\mathbf{MB}]$ ,	$N = [\mathbf{MB}]$ ,	$N = [\mathbf{MH}]$ ,
where $M = \frac{1}{c} IS$	where $M = IS$	where $\mathbf{M} = m\mathbf{l}$

Formulas relating fields to their sources:

Field due to current		Field due to magnet	
CGS	SI	CGS	SI
$d\mathbf{H} = \frac{I}{cr^2} \left[ d\mathbf{l}, \frac{\mathbf{r}}{r} \right]$	$d\mathbf{H} = \frac{I}{4\pi r^2} \left[ d\mathbf{l}, \frac{\mathbf{r}}{r} \right]$	$B = \frac{m}{r^2}$	$B = \frac{m}{4\pi r^2}$
$B = \mu H$	$B = \mu_0 \mu H$	$B = \frac{2M}{r^3}$	$B = \frac{M}{2\pi r^3}$

Substituting any of the lower formulas in any of the upper ones and using the relation  $\mathbf{B} = \mu_0 \mu \mathbf{H}$ , we obtain the formulas for magnetic, electromagnetic, magnetoelectric and electrodynamic interaction. We shall illustrate each type of interaction by an example.

*Magnetic interaction*, i.e., the action of one magnet on another. Two poles separated by a distance  $r$  interact in accordance with Coulomb's law, i.e.,

$$F = \frac{m_1 m_2}{\mu r^2} \text{ (CGS)} \quad \text{or} \quad F = \frac{m_1 m_2}{4\pi \mu_0 \mu r^2} \text{ (SI)}.$$

The interaction force is inversely proportional to the magnetic permeability.

*Electromagnetic action*, i.e., the action of a current on a magnet. A current element exerts a torque on a magnetic needle. For simplicity, we assume that  $\mathbf{M} \perp \mathbf{H}$ , i.e., the magnetic needle is perpendicular to the flux lines. Then,

$$dN = \frac{MI}{cr^2} dl \sin \angle d\mathbf{l}, \mathbf{r} \text{ (CGS)}$$

or

$$dN = \frac{MI}{4\pi r^2} dl \sin \angle d\mathbf{l}, \mathbf{r} \text{ (SI)}.$$

The interaction does not depend on the magnetic permeability, i.e., on the properties of the medium.

*Magnetoelectric action*, i.e., the action of a magnet on a current. Consider a current-carrying circuit located along the extension of the bar magnet axis at



a distance  $r$  from the magnet (Fig. 113). The torque acting on the circuit is

$$N = M_{cur} B \sin \alpha = \frac{M_{cur} M_{mag}}{r^3} \sin \alpha \text{ (CGS)}$$

or

$$N = \frac{M_{cur} M_{mag}}{2\pi r^3} \sin \alpha \text{ (SI)}.$$

Thus, the interaction does not depend on the magnetic permeability.

*Example.* A circuit of area  $S = 20 \text{ cm}^2$ , through which a current  $I = 10 \text{ A}$  flows, interacts at a distance of 100 cm with a bar magnet whose magnetic moment is  $M_{mag} = 1,000 \text{ CGS units} = 1 \text{ A} \times \text{m}$ . The torque acting on the circuit is

$$N = \frac{2M_{cur} M_{mag}}{r^3}.$$

$$M_{cur} = \frac{1}{3 \times 10^{10}} \times 10 \times 20 = \frac{2}{3} \times 10^{-8} \text{ CGS unit}$$

$$N = 4 \times 10^{-5} \text{ dyne} \times \text{cm} = 0.04 \text{ N} \times \text{m}.$$

*Electrodynamic action*, i.e., the action of one current on another current. Two parallel currents are attracted with a force

$$dF = \frac{I_1}{c} dl_1 B,$$

i.e.,

$$dF = \frac{I_1 I_2 dl_1 dl_2}{c^2 r^2} \text{ (CGS)} \quad \text{or} \quad dF = \mu_0 \mu \frac{I_1 I_2 dl_1 dl_2}{4\pi r^2} \text{ (SI)}.$$

The interaction is directly proportional to the magnetic permeability.

The formulas for the interaction of magnetic systems may be derived in exactly the same manner.

*Example.* It is essential to take into account electrodynamic interaction in the laying of bus bars. If a short circuit should occur, the bus bars and their supporting insulators must be sufficiently firm to withstand large electrodynamic forces. Assume that a current  $I_1 = I_2 = 3 \times 10^4 \text{ A}$  flows in parallel bus bars separated by a distance  $d = 20 \text{ cm}$ . A force  $F = BI = \mu_0 HI$  acts on a unit length of each bus bar, where  $H = \frac{I}{2\pi d}$  is the magnetic field intensity created by the linear current flowing in the other bus bar. Thus,

$$F = \frac{\mu_0 I^2}{2\pi d} = \frac{4\pi \times 10^{-7} \times 9 \times 10^8}{2\pi \times 0.2} = 900 \text{ N},$$

i.e., a force of  $\sim 90 \text{ kgf}$  acts on each metre of a bus bar. The same result could be obtained by integrating the last formula for  $dF$  above.

### Sec. 103. EQUIVALENCE OF CURRENTS AND MAGNETS

We have called attention to the fact that a similarity exists between the expressions for the torque acting on a magnetic needle and the torque acting on a current-carrying circuit. As a matter of fact, these two systems behave extremely alike in an external field. If each of the systems is characterised by its magnetic moment vector, the likeness appears even greater. Both systems tend to become oriented in the magnetic field with the magnetic moments parallel to the flux lines of the

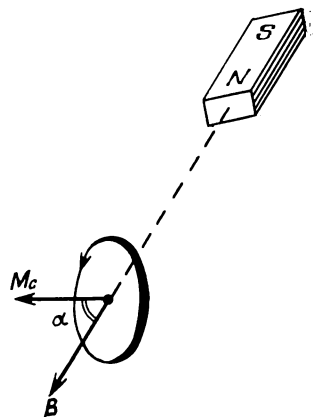


Fig. 113

field. If the magnetic moment is displaced from the position of stable equilibrium, a torque  $N = [MH]$  acts on the system in the case of the magnetic needle and a torque  $N = [MB]$  in the case of a current-carrying circuit. The potential energies of these systems are represented by the formulas  $U = -MH$  and  $U = -MB$ , respectively.

If we recall that  $B = \mu_0 \mu H$ , the difference between the two formulas becomes immediately evident, i.e., they differ only with respect to the magnetic permeability factor. Hence, as regards mechanical action, a magnetic needle of moment  $M$  is equivalent to a current-carrying circuit of moment  $M_{cur} = \frac{M}{\mu_0 \mu}$ .

However, the analogy between these two systems goes even further. We shall now show that the proper fields of a magnetic needle and a current-carrying circuit are alike except for a constant factor. This similarity occurs at distances considerably greater than the dimensions of the system. We shall prove this for a point in space in line with the magnetic moment at a distance  $r$  from the centre of the system. The field of a magnet for such a point has already been calculated and found to be equal to  $B = \frac{2M}{r^3}$ . It remains to determine the field of a circular loop of current along the axis of the loop.

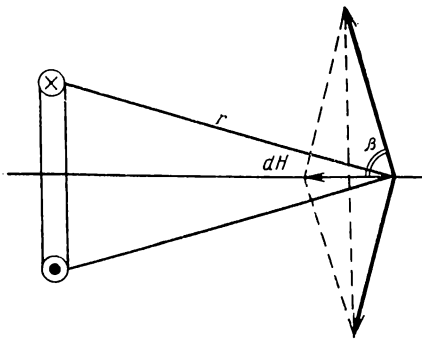


Fig. 114

In Fig. 114 are shown the vectors of the field intensities due to two arc elements of current directed into and out of the page, respectively. The field intensity vectors are perpendicular to their corresponding

current elements and radius vectors, i.e., they are in the plane of the page. The sense of the intensity vectors is determined from the vector product rule or, what amounts to the same, the right-hand screw rule.

Since a current element and its radius vector are perpendicular to each other, the field created by such an element is equal in this case to  $dH = \frac{Idl}{cr^2}$ . Now, let us add the two vectors depicted in the figure. For the field created by the "antipodes", we obtain

$$dH = \frac{2I dl}{cr^2} \cos \beta,$$

where the meaning of the various designations is evident from the figure. This value of field is yielded by every such pair of "antipodes". Therefore, the resultant field is obtained from the last expression by replacing  $dl$ , the length of an element, by  $\pi a$ , half the circumference of a circle. Thus, the field intensity along the axis of a circular current, at a distance  $r$  from the current,\* is given by the formula

$$H = \frac{2\pi a^2 I}{cr^3}.$$

Since  $\frac{1}{c} IS$  is the moment of a circular current (here  $S = \pi a^2$ ), we obtain:  $H = \frac{2M}{r^3}$  and  $B = \mu_0 \mu \frac{2M}{r^3}$ .

\* Since we are dealing with large distances, the difference between  $r$  and the distance to the centre of the system is negligible.

Thus, we have proved that a magnetic dipole and a current-carrying circuit are not only equivalent as regards the forces acting on them, but also as regards the fields created by them. The nature of the equivalence is the same in both cases. A magnetic needle of moment  $M$  can be replaced by a current-carrying circuit of moment

$$M_{\text{cur}} = \frac{M}{\mu\mu_0}.$$

In vacuum and for the CGS system, i.e., for  $\mu\mu_0 = 1$ , the equivalence principle becomes even simpler. In this case, a magnetic needle of moment  $M$  is equivalent to a current-carrying circuit having the same moment.

*Examples.* 1. Let us return to the example on p. 204 and calculate the magnetic induction of the same magnet in the practical system of units:

$B = 0.05$  T,  $S = 3 \times 10^{-4}$  m<sup>2</sup>,  $\Phi = 15 \times 10^{-6}$  V sec,  $m = 15 \times 10^{-6}$  V sec, and  $l = 0.1$  m, and  $M = ml = 15 \times 10^{-6} \times 0.1 = 1.5 \times 10^{-6}$  V sec metre. Hence,

$$B = \frac{M}{2\pi r^3} = 2.4 \times 10^{-7} \text{ T},$$

which is in complete agreement with the result obtained on p. 204.

2. Consider a current-carrying circuit for which  $I = 5$  A and  $S = 2$  cm<sup>2</sup>. The magnetic field intensity created at a distance  $r = 50$  cm (along the axis of the circuit and perpendicular to its plane) is  $H = \frac{2M}{r^3}$

$$M = \frac{1}{c} IS = \frac{1}{3 \times 10^{10}} 5 \times 3 \times 10^9 \times 2 = 1 \text{ erg/G}, \quad \text{and} \quad H = 1.6 \times 10^{-5} \text{ oersted (Oe)}.$$

#### Sec. 104. ROTATIONAL NATURE OF A MAGNETIC FIELD

A study of the nature of magnetic lines shows that magnetic lines differ basically from electric fields. Electric lines have a beginning and an end, i.e., there are no closed lines in a constant electric field. On the other hand, experiments show that magnetic flux lines, i.e., vector lines of magnetic induction, are always closed. In other words, such lines have neither beginning nor end.

For reasons discussed above, force fields in which the work along a closed path is equal to zero are known as potential fields. Vector fields characterised by closed flux lines are known as rotational fields. A magnetic field is a rotational field.

If we describe a closed surface in a magnetic field, the net outward flux  $\Phi = \oint B \cos \alpha dS$  through such a surface will always be equal to zero. In other words, the number of lines entering this surface is equal to the number of lines leaving it. Thus, the equation  $\oint B \cos \alpha dS = 0$  is the mathematical expression of the fact that magnetic flux lines have neither beginning nor end.

The magnetic lines always encircle the currents creating the field. Therefore, the integrals taken along induction or intensity flux lines, i.e.,  $\oint B dl$  and  $\oint H dl$ , respectively, differ from zero. It is more convenient to consider the second integral, for its value is proportional to the magnitude of the electric current encircled by flux lines. This can be seen from the basic field intensity formula, which shows that  $H$  and current strength are directly proportional to each other.

By analogy with electrostatics,  $\int H dl$  is called the magnetomotive force (mmf).

If the integral is taken along a flux line, then

$$\int H \, dl = \int H \, d\mathbf{l}.$$

The magnetomotive force along a closed curve is proportional to the current encircled by this curve:

$$\oint H \, dl = kI,$$

where  $k$  is a coefficient of proportionality.

A flux line may encircle more than a single current. Then, using the algebraic sum of the currents, the equation assumes the form

$$\oint H \, dl = k \sum I.$$

Deeper theoretical analysis, which we are unable to go into here, shows that the above equation is subject to two more generalisations. First, the integral need not be taken along a flux line, but can be taken along any arbitrary circuit. Secondly, the coefficient of proportionality in the equation is a constant depending only on the properties of the medium and is the same for all geometric conditions. Thus, the magnetomotive force is the same for any closed curve that encircles a current of specified strength. The shape of the curve and its length are of no significance. It is immaterial whether the curve encircles one current or ten currents and whether these currents are rectilinear or curved. As long as the algebraic sum of the currents passing through the closed curve remains the same, so does the magnetomotive force.

Since the coefficient of proportionality in the formula for the magnetomotive force is a universal constant, we can determine  $k$  if the magnetomotive force can be calculated for any system whose field is known.

We are familiar with the general expression for the magnetic field intensity of an elementary current, but mathematical difficulties are encountered in calculating the magnetomotive force by means of the formula for the field intensity, namely,

$$d\mathbf{H} = \frac{I}{cr^2} \left[ d\mathbf{l}, \frac{\mathbf{r}}{r} \right].$$

However, we are also familiar with the formula for the magnetic field intensity along the axis of a circular current, namely,  $H = \frac{2M}{r^3}$ . No special difficulties are encountered in calculating the magnetomotive force along this axis. We should not be disturbed by the fact that the integration is carried out along a straight line, while we are interested in the magnetomotive force along a closed curve. As a matter of fact, a straight line extending from minus infinity to plus infinity is a closed curve, for it is closed at infinity. The expression for the magnetomotive force,  $\int H \, dl$ , along such a closed curve, i.e., along the axis of a circular current from minus infinity to plus infinity, may be written in the form

$$2M \int_{-\infty}^{+\infty} \frac{dl}{(\sqrt{l^2 + a^2})^3},$$

where  $a$  is the radius and  $l$  is the distance measured along the axis of the circuit. The integral is easily evaluated by using the new variable  $\beta$ , defined by the formula  $\frac{l}{a} = \cot \beta$ . It turns out to be equal to  $\frac{2}{a^2}$ . Substituting  $\frac{1}{c} I \pi a^2$  for  $M$  and equating

the value of the magnetomotive force to  $kI$ , we obtain

$$k = \frac{4\pi}{c} \quad (\text{in the CGS system})$$

$$k = 1 \quad (\text{in the SI system}).$$

The magnetomotive force relation now assumes the form

$$\oint \mathbf{H} \, d\mathbf{l} = \frac{4\pi}{c} \sum I \quad \text{or} \quad \oint \mathbf{H} \, d\mathbf{l} = \sum I.$$

The magnetomotive force relation is very useful in determining the magnetic fields of various systems. Its application is facilitated by considerations of symmetry and in this respect the following discussion is quite analogous to the discussion relating to the solution of the corresponding problems in electrostatics by means of the Gauss-Ostrogradsky theorem.

Let us first consider an infinitely long rectilinear current. From considerations of symmetry it is evident that the flux lines must be circles whose centres lie on the wire axis. It is similarly evident that at all points on such a circle the numerical value of the intensity is the same. Applying the magnetomotive force relation to such a flux line, we obtain:  $H \oint d\mathbf{l} = \frac{4\pi}{c} I$ . Here,  $\oint d\mathbf{l}$  is simply the length of a flux line. If the points under consideration are located at a distance  $r$  from the wire axis, then  $\oint d\mathbf{l} = 2\pi r$ . Thus, for the magnetic field of an infinitely long rectilinear current in the region outside the wire, we obtain

$$H = \frac{2I}{cr} \quad (\text{in the CGS system}).$$

$$H = \frac{I}{2\pi r} \quad (\text{in the SI system}).$$

Let us now determine the magnetic field intensity inside the wire. Assume that the radius of the wire is designated by  $a$  and that the current density is uniform over the wire cross-section. Hence, the flux lines within the wire must also be circular. Consider such a flux line of radius  $r$ . The current flowing through it will be  $\frac{r^2}{a^2} I$ . Therefore, the magnetomotive force relation yields:

$$H \times 2\pi r = \frac{4\pi}{c} \frac{r^2}{a^2} I, \quad \text{and} \quad H = \frac{2}{c} \frac{r}{a^2} I.$$

In the SI system

$$H = r \frac{I}{2\pi a^2}.$$

Thus, we see that the magnetic field intensity along the wire axis is equal to zero. The intensity increases with radius and becomes a maximum at the surface of the wire. Then, for the region outside the wire, the field intensity decreases, being inversely proportional to the distance from the axis (Fig. 145).

If the field is determined at a point for which the distance  $r$  is much less than the distance to the end of a wire, then the formula  $H = \frac{I}{2\pi r}$  is also valid for a wire of finite dimensions.

*Example.* Let us calculate the magnetic field intensity at a distance of 5 cm from the axis of a rectilinear current of 20 A.

In the CGS system ( $I = 20 \times 3 \times 10^9 = 6 \times 10^{10}$  CGS units);

$$H = \frac{2I}{cr} = \frac{2 \times 6 \times 10^{10}}{3 \times 10^{10} \times 5} \times 0.8 \text{ Oe.}$$

In the SI system ( $I = 20$  A and  $r = 0.05$  m):

$$H = \frac{I}{2\pi r} = \frac{20}{2\pi \times 0.05} = 64 \text{ A/m.}$$

Another important example of the application of the magnetomotive force relation is in the calculation of the field of a solenoid.

Consider a uniformly wound toroidal solenoid whose circumferential length is  $L$ . The field within the solenoid is uniform and all the flux lines are concentric with  $L$ . This system plays the same role in magnetic field theory as a parallel-plate condenser of infinite extent in electric field theory. Each flux line envelops all  $n$  turns. Therefore, the magnetomotive force along a flux line of length  $L$  is

$$\oint H \, dl = \frac{4\pi}{c} nI.$$

Since  $\oint H \, dl = HL$ , we obtain

$$H = \frac{4\pi}{c} \frac{n}{L} I \text{ (CGS)}$$

$$H = \frac{n}{L} I \text{ (SI).}$$

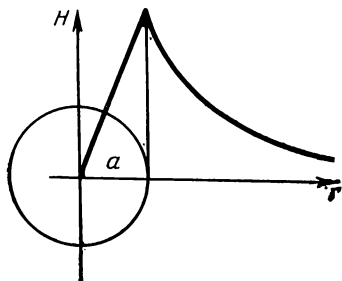


Fig. 115

The magnetic field intensity of the coil is determined by its "ampere turns", i.e., the product of the current strength and the number of turns per unit length of the solenoid. It should be noted that the simplicity of the last formula is one of the justifications for the electrical engineering system of expressing the field equations. Since a solenoid is one of the basic elements of electrical engineering devices, simplification of the formula for the calculation of its magnetic field intensity is of great practical significance.

The formula  $H = \frac{n}{L} I$  is also valid for a straight solenoid if used to determine the field within the solenoid at points sufficiently far away from the edges.

*Example.* The magnetic field intensity at the centre of a long, thin solenoid, where  $L = 15$  cm,  $n = 1,500$  turns and  $I = 0.1$  A, is

$$H = 1,000 \text{ A/m.}$$

In the CGS system:

$$H = \frac{4\pi}{c} \frac{n}{L} I = \frac{4\pi}{3 \times 10^{10}} \frac{1,500}{15} (0.1 \times 3 \times 10^9) = 12.56 \text{ Oe,}$$

$$1 \text{ A/m} = 4\pi \times 10^{-3} \text{ Oe} \quad \text{and} \quad 1 \text{ Oe} \approx 80 \text{ A/m.}$$

#### Sec. 105. LAW OF ELECTROMAGNETIC INDUCTION AND LORENTZ FORCE

The phenomenon of electromagnetic induction discovered by Faraday, the great English physicist, may be described as follows: an electric current is induced in a closed conductor loop if the value of the magnetic flux passing through the loop changes. Moreover, the induced emf is proportional to the rate of change of the magnetic flux, i.e., the derivative with respect to time

$$\frac{d\Phi}{dt}, \quad \text{where} \quad \Phi = \int B \cos \alpha \, dS.$$

We shall show that the law of electromagnetic induction is closely related to the existence of a Lorentz force. If the electromagnetic induction is due to the displacement of a wire in a magnetic field, then the law of induction follows from the expression for the Lorentz force.

In order to avoid confusion due to difficulties of a purely mathematical nature, let us simplify the proof by assuming that the induced emf arises in a rectangular circuit oriented perpendicular to the flux lines in a uniform magnetic field. A change in flux produces a translatory displacement of one side of the rectangle of length  $l$  as shown in Fig. 116. Since there are free charges in the displaced conductor, these charges are subjected to the action of a

Lorentz force  $f = \frac{e}{c} vB$  when the conductor moves

with a velocity  $v$ . In view of the fact that the velocity, magnetic field and conductor are at right angles to each other, we are able to dispense with the vector notation in the formula for the force since the sine of the angle is equal to unity. The Lorentz force is directed perpendicular to the plane passing through the direction of the velocity  $v$  of the charges, and hence of the wire, and the direction of the magnetic flux lines. Thus, the force is directed along the wire. The charges are impelled to move and an induced current is thereby created.

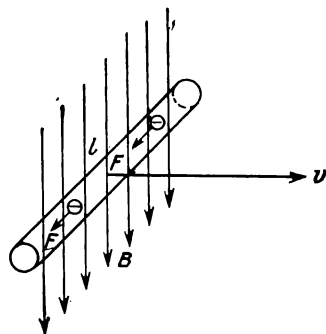


Fig. 116

The *electromotive force* is equal to the work of moving a unit charge around a closed circuit. Since the force acting on a unit charge is equal to  $\frac{1}{c} vB$ , the work of this force along the moving wire is equal to  $\frac{1}{c} vBl$ . Moreover, no work is performed along the rest of the circuit. Hence, the last expression sought for the induced emf. It has the form:

$$\mathcal{E}^{ind} = \frac{1}{c} vBl \text{ (CGS)}$$

$$\mathcal{E}^{ind} = vBl \text{ (SI)}$$

Assume that the wire moves a distance  $dx$  during the time  $dt$ . Thus, the area of the circuit increases by the amount  $l dx = dS$  and the magnetic flux by the amount  $d\Phi = BdS$ . Since  $v = \frac{dx}{dt}$ , the induced emf may also be written in the following form:  $\frac{1}{c} \frac{BdS}{dt}$ . But this is precisely the expression for Faraday's law of electromagnetic induction, i.e.,  $\mathcal{E}^{ind} = -\frac{1}{c} \frac{d\Phi}{dt}$  in the CGS system and  $\mathcal{E}^{ind} = -\frac{d\Phi}{dt}$  in the SI system.

We have thus demonstrated that electromagnetic induction and the deflection of moving electric charges in an external field represent different aspects of one and the same law of nature. In the next chapter, we shall return to this interesting problem. Here, it was only necessary to present the essence of the electromagnetic induction law.

#### Sec. 106. MEASUREMENT OF MAGNETIC FIELDS BY MEANS OF INDUCED IMPULSES

The phenomenon of electromagnetic induction is utilised in the design of precision instruments for the measurement of magnetic fields. Assume that it is necessary to determine the value of the magnetic field at some point in space. A small,

flat coil or a single turn of wire is placed in a magnetic field perpendicular to the flux lines and the ends of the winding are connected by means of leads to the terminals of a ballistic galvanometer. Now, if the coil is rapidly turned through a  $90^\circ$  angle in such a manner that its flat surface becomes parallel to the flow lines, an electrically induced current flows in the winding as the coil is being turned. The brief flow of current, which rapidly reaches a maximum and then drops to zero, is called an *induced impulse* (Fig. 117). During this brief interval, a certain quantity of electricity flows in the wire. The charge can be very accurately measured by means of a ballistic galvanometer, a device having a moving coil with a high moment of inertia that integrates the electric current over the period of the impulse.

If the resistance of the coil is  $R$  and the number of turns  $n$ , the induced current strength may be written in the form

$$I = \frac{n}{R} \frac{d\Phi}{dt}.$$

The quantity of electricity flowing in the wire during the period of the induced impulse is

$$Q = \int_0^{\tau} I dt = R^{-1}n \int_1^2 d\Phi = R^{-1}n (\Phi_2 - \Phi_1),$$

where  $\Phi_1$  is the value of the flux passing through the coil in the first position and  $\Phi_2$  is the value in the second position.

If  $\Phi_1$  or  $\Phi_2$  is equal to zero, i.e., magnetic flux lines do not pass through the coil in the initial or final position, the performed measurement yields the value of the magnetic induction. For this purpose, we need only divide the value of the magnetic flux by the coil area  $S$ .

$$\text{i.e., } B = \frac{QR}{nS}$$



Fig. 117

Of course, other methods of measurement are also possible. Thus, instead of rotating the coil, the field may be switched on or off. Also, to make the effect more pronounced, the coil can be turned through an angle of  $180^\circ$  instead of  $90^\circ$ . This doubles the effect. Similarly the polarity of the field can be reversed, rather than simply switched on or off.

Since the measuring coil may be made as small as a square millimetre, measurements by this method enable us to accurately determine the magnetic field in small regions.

This method is also used to measure magnetomotive force. For this purpose, we use a measuring device known as a Rogovsky belt—a long coil on a flexible belt. The belt may be shaped into any desired form and its two ends placed at any two points in a given region. Also, if desired, the ends of the belt may be brought into contact with each other. We shall show that when the field is switched off the deflection of a ballistic galvanometer connected to such a measuring belt is proportional to the magnetomotive force along the path of the flexible belt.

The deflection of the ballistic galvanometer is a measure of the quantity of magnetic flux passing through all the turns of the coil. Let  $n$  be the winding density, i.e., the number of turns per unit length of the measuring belt. Then, on a small belt segment  $\Delta l_i$ , there are  $n\Delta l_i$  turns, and the magnetic flux passing through these  $n\Delta l_i$  turns is equal to  $\Phi_i n \Delta l_i$ .

If the medium is homogeneous and all the turns have the same area, then

$$\Phi_i = \mu S H_i,$$



and the total magnetic flux passing through the entire measuring belt is

$$\Phi = \sum_i \mu S n H_i \Delta l_i.$$

Taking the limit for  $\Delta l_i \rightarrow 0$ , we obtain

$$\Phi = \mu S n \int_1^2 H dl.$$

Since the measurement occurs in a medium for which  $\mu$  does not significantly differ from  $\mu_0$ , the quantity  $\mu S n$  is a constant of the instrument. The throw of the ballistic galvanometer in these belt measurements is exactly proportional to the magnetomotive force between the points at which the ends of the belt are located.

By means of this device, it is easy to demonstrate the validity of the laws discussed in Sec. 104. Thus, as long as the coil encircles one and the same current, the magnetomotive force will remain the same for all configurations. Also, it is easily verified that the magnetomotive force along a circuit not encircling currents is equal to zero. For the case when the coil encircles a current several times, the magnetomotive force can be shown to increase by the corresponding number of times, etc.

It should be emphasised that magnetic field measurements by means of induced impulses are of particular importance when we are concerned with the magnetic field inside a solid body. The only other method available is to make slit in the solid body. Such a procedure is usually not possible.

Let us consider the most common problem—the measurement of the magnetic permeability of an iron body. The most accurate results are obtained when the substance under investigation is in the form of a toroid. Two windings are wound on such a ring—one connected to a current source and the other to a ballistic galvanometer. If current is flowing, a magnetic flux  $\Phi = BS$  passes through the ring. By reversing the direction of the current through the primary winding, an induced current is produced in the secondary. The quantity of electricity  $Q$  flowing through the galvanometer is related to the magnetic induction inside the ring by a relation already discussed, namely:

$$B = \frac{QR}{n_2 S},$$

where  $S$  is the cross-section of the toroid (assuming the turns are wound tight on the surface of the ring),  $n_2$  is the number of secondary turns and  $R$  is the resistance of the secondary winding. As regards the magnetic field intensity, it may be determined by the formula for a ring solenoid, namely,  $H = \frac{n_1 I}{L}$ . The magnetic permeability of the substance under investigation is then equal to  $B$  divided by  $H$ .

#### Sec. 107. FINITE BODIES IN A MAGNETIC FIELD

To one degree or another all bodies possess magnetic properties. These are indicated, first, by the fact that a magnetic field exerts forces and torques on bodies and, secondly, by the fact that a body placed in a magnetic field distorts the field. As was indicated above, the magnetic properties of a substance are characterised by the coefficient  $\mu$ —the magnetic permeability of the substance. In accordance with the value of  $\mu$ , bodies may be divided into three distinct classes: ferromagnetic substances—including iron, nickel and cobalt—whose relative perme-

abilities are much greater than unity; paramagnetic substances, whose permeabilities are somewhat greater than unity; and diamagnetic substances, whose permeabilities are slightly less than unity. Typical values are given in the table.

Substance	$\mu$	$\chi$
Copper . . . . .	0.999990	$-10^{-5}$
Water . . . . .	0.999991	$-9 \times 10^{-6}$
Platinum . . . . .	1.000300	$300 \times 10^{-6}$
Silicon . . . . .	0.999986	$-14 \times 10^{-6}$
Tungsten . . . . .	1.000079	$79 \times 10^{-6}$

When a diamagnetic or paramagnetic body is placed in a magnetic field, the distortion of the field is negligible. On the other hand, when a ferromagnetic body is placed in the field, there is considerable distortion.

The forces exerted by magnetic fields may be detected without particular difficulty in the case of paramagnetic and diamagnetic bodies. In the case of iron objects everybody is familiar with the fact that magnetic fields exert large forces.

Let us first consider magnetic forces. A body that does not possess magnetic properties becomes magnetic when placed in a field. This magnetisation process manifests itself in the acquisition of a magnetic moment by the body. As we know, a system possessing a magnetic moment may be detected in two ways. In a uniform field, the body tends to become oriented in such a manner that the direction of the moment is parallel to the external field. Moreover, in a nonuniform field the body will experience a force tending to move it along the lines of force.

In the case of ferromagnetic bodies, the torque may be detected without difficulty. The magnetic moment of the body may then be determined from the formula  $N = [MH]$ . However, we are not usually interested in a body having a particular shape, but are interested rather in the substance as such. Therefore, when possible the measured value is recalculated on the basis of unit volume. The vector directed along the magnetic moment and numerically equal to the magnitude of the magnetic moment per unit volume is called the *magnetisation vector*  $J$ . Of course, the magnetisation vector can be determined without difficulty from the magnetic moment of a body only if we are sure that the magnetisation of the sample is uniform. This is the case when the sample is in the form of an ellipsoid or a degenerate ellipsoid, i.e., a cylinder, plate or sphere (see p. 196). That is why such bodies are used in experiments of this kind.

Determination of the magnetisation vector by measuring the torque is easily accomplished for ferromagnetic bodies. Since the torque is very small for paramagnetic and diamagnetic bodies, this measurement is difficult to perform. It is therefore preferable in these cases to measure the forces acting on a body located in a nonuniform field.

Let us consider a small volume of a magnetic substance located in a nonuniform field. For simplicity, assume that the field varies along one axis and that the gradient is equal to  $\frac{dH}{dx}$ . Since a small volume of magnetic substance behaves like a magnetic dipole, the potential energy of a unit volume may be written in the form  $U = -JH$ . If the moment acts along the field, the force exerted on a unit volume of the magnetic substance is equal to the derivative of the potential ener-

gy with respect to the coordinate, i.e.,

$$f = -J \frac{dH}{dx}.$$

Thus, if we know the field gradient, we can determine the magnetic moment of a unit volume of the body under investigation by measuring the force. In practice, there are various ways of accomplishing this, the simplest being by means of a so-called magnetic balance. A thread is passed through an aperture made in one of the pans of an analytical microbalance. Then, the sample is attached to the end of the thread and suspended between the poles of a magnet. Before and after energising the magnet, the sample is balanced; hence, the difference between the readings is equal to the value of the force  $f$ .

The weights must be quite accurate as can be seen from the following example. A piece of bismuth, a substance whose diamagnetic properties are most pronounced, has a magnetisation  $J$  of  $2 \times 10^{-2}$  CGS unit when placed in a magnetic field whose intensity  $H$  is  $\sim 1,000$  Oe. If the nonuniformity of the magnetic field is  $\frac{dH}{dx} \sim 50$  Oe/cm, a force of only 1 dyne will be exerted on each cubic centimetre of bismuth, i.e.,  $f \sim 1$  dyne/cm<sup>3</sup>.

Experiments show that for diamagnetic and paramagnetic bodies the following simple relationship exists between the magnetisation vector and the magnetic field intensity:

$$J = \mu_0 \kappa H,$$

where  $\kappa$  is the magnetic susceptibility. For diamagnetic bodies  $\kappa$  is negative, while for paramagnetic bodies it is positive. Values of  $\kappa$  are given in the table on p. 216. When  $\kappa$  is positive the magnetisation vector is in the direction of the field intensity vector, but when  $\kappa$  is negative, i.e., for diamagnetic bodies, the magnetisation vector is opposite to the field.

Due to this difference in sign, the behaviours of these two classes of bodies under identical conditions are completely different. This is illustrated in Fig. 118. As can be seen, the differences are quite striking. A paramagnetic body is attracted

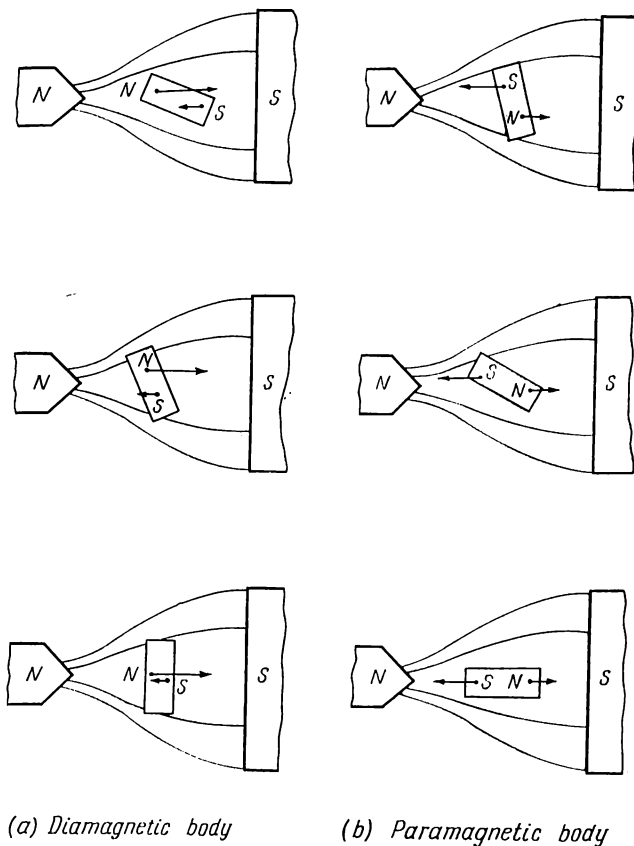


Fig. 118

toward the region of strong field, while a diamagnetic body is repelled from such a region. In a uniform field, a paramagnetic needle tends to become oriented with its axis along the flux lines, while a diamagnetic needle tends to become oriented perpendicular to the flux lines (see the analogous example in the case of a dielectric, p. 195).

The determination of magnetic susceptibility by measuring the force in a nonuniform field may be accomplished for solid bodies in the form of a monocrystal or powder. This method is also easily adapted to liquids. In this case, the experiment can be so arranged that the measured quantity is the increase or decrease in the level of the liquid as it is attracted toward, or repelled from, the region between the poles of a magnet.

#### Sec. 108. RELATIONSHIP BETWEEN PERMEABILITY AND SUSCEPTIBILITY

Both permeability and susceptibility may be determined straightforwardly. The permeability is determined from the formula  $\mu = \frac{B}{H}$  by measuring the induction and field intensity. The susceptibility is determined, as described in the preceding article, from the forces exerted on a magnetic substance. To be sure, the relationship between these two characteristics of the magnetic properties of a substance can be established experimentally. However, there is no need to do this since an exact and simple relationship exists between  $\mu$  and  $\kappa$ . This will now be shown.

Let us return to the experiment for determining the magnetic permeability of a body in the form of a toroid. The primary winding creates a field of intensity  $H = \frac{nI}{L}$ , which is independent of the substance of the toroid, i.e., if the toroid were not present, the field intensity would be given by the same formula. The situation is different as regards the magnetic flux. It can be shown experimentally that the value of  $B$  depends on the magnetic permeability. If an iron core is placed in the coil,  $B$  becomes hundreds or thousands of times greater than in the case of an air core. This increase in magnetic flux is due to the magnetisation effect.

First, it should be noted that the magnetic induction of a ring solenoid without an iron core ( $\mu_0 H$ ) has the significance of magnetic moment per unit volume.

The magnetic moment of one turn is equal to  $IS$ , for in this discussion we shall employ the SI system. For the total magnetic moment of the system we obtain  $nIS$ , and the magnetic moment in unit volume,  $\frac{nIS}{LS}$ , is simply equal to the field intensity. The magnetic moment of the equivalent dipoles is  $\mu_0$  times greater (cf. Sec. 103). Therefore, the magnetic induction  $\mu_0 H$  of a uniform magnetic field created by the turns of a ring solenoid without an iron core can be expressed as the magnetic moment of the equivalent dipoles per unit volume.

We are entirely justified in assuming that the significance of magnetic induction is maintained if, without disturbing the uniformity of the field, the coil is filled with an additional number of magnetic dipoles. If  $J$  is the magnetic moment per unit volume due to the additional dipoles, the magnetic induction increases by this amount and becomes

$$B = \mu_0 H + J.$$

Such an increase in  $B$  also occurs when the solenoid is filled with a magnetic substance. Since  $J = \mu_0 \kappa H$ , then  $B = \mu_0 (\kappa + 1) H$ ; hence, the susceptibility and permeability are related by the equation

$$\mu = 1 + \kappa.$$

Analogous calculations employing the Gaussian system of units lead to formulas with other coefficients. The magnetic moment of currents (and dipoles) per unit volume is

$$\frac{n \frac{1}{c} IS}{LS} = \frac{M}{4\pi} H.$$

Therefore, in the presence of a medium,

$$\frac{1}{4\pi} B = \frac{1}{4\pi} H + J, \quad \text{i.e.,} \quad B = H + 4\pi J.$$

Letting  $J = \kappa' H$ , we obtain

$$B = (1 + 4\pi\kappa') H.$$

Hence,

$$\mu = 1 + 4\pi\kappa', \quad \text{where} \quad \kappa' = \frac{\kappa}{4\pi}.$$

*Example.* Let us perform the calculation in the example on p. 217 employing the SI system. For bismuth,  $\kappa' = 2 \times 10^{-6}$ , i.e.,  $\kappa = 4\pi\kappa' = 8\pi \times 10^{-6}$ . The piece of bismuth is in a magnetic field whose intensity is

$$H = 1,000 \text{ Oe} = \frac{1}{4\pi} \times 10^3 \times 1,000 \frac{\text{A}}{\text{m}} = \frac{10^6}{4\pi} \frac{\text{A}}{\text{m}}.$$

Furthermore, the nonuniformity is expressed by

$$\frac{dH}{dx} = 50 \frac{\text{Oe}}{\text{cm}} = 50 \times \frac{1}{4\pi} \times 10^3 \times 100 \frac{\text{A}}{\text{m}^2} = \frac{5 \times 10^6}{4\pi} \frac{\text{A}}{\text{m}^2}.$$

The bismuth magnetisation is given by  $J = \mu_0 \kappa' H = 8\pi \times 10^{-6} \frac{\text{V sec}}{\text{m}^2}$ . Hence, the force acting on a unit volume (1 metre<sup>3</sup>) is

$$f = J \frac{dH}{dx} = 8\pi \times 10^{-6} \times \frac{5 \times 10^{-6}}{4\pi} = 10 \frac{\text{N}}{\text{m}^3}.$$

Clearly,  $10 \frac{\text{N}}{\text{m}^3} = 1 \frac{\text{dyne}}{\text{cm}^3}$ , which agrees with the result obtained in the previous example.

#### Sec. 109. DISTORTION OF A MAGNETIC FIELD DUE TO THE PRESENCE OF A MAGNETIC SUBSTANCE

The problem of magnetic field distortion has practical significance only when the distortion is due to an iron body. To a large extent, we shall be repeating the analysis given on p. 196 for the analogous case of dielectric bodies.

At the boundary of two media of different magnetic permeability, the magnetic field vectors (induction as well as intensity) are refracted. To determine the law of refraction, let us first consider the magnetomotive force along a small circuit  $ABCD$  whose sides parallel to the boundary surface are close to each other, but on either side of the boundary as shown in Fig. 119. Since no current flows through this circuit, the magnetomotive force is equal to zero. Let us resolve the magnetic field intensity vectors on either side of the boundary into normal and tangential components. From the figure it is evident that the magnetomotive force can be equal to zero only if the tangential components are equal to each other:

$$H_{1t} = H_{2t}.$$

Another condition at the boundary between two media is established by considering the magnetic flux passing through a small cylinder (not shown in the figure) enclosing a portion of the boundary. Since the magnetic lines have no sources, the number of flux lines entering the cylinder through the top is equal to the number

leaving through the base. The lateral surface has an infinitely small area and the flux through it is equal to zero. Now, let us resolve the magnetic induction vectors on either side of the boundary into normal and tangential components. Clearly,

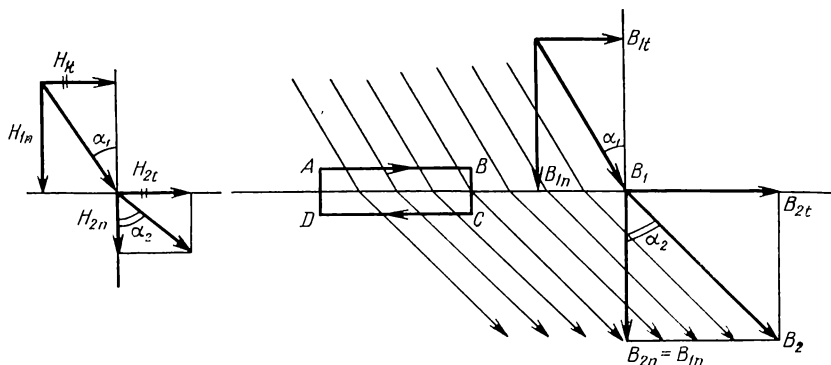


Fig. 119

the flux entering the cylinder can be equal to the flux leaving the cylinder only if the normal components of the induction vectors do not change in crossing the boundary:  $B_{1n} = B_{2n}$ .

From these two rules, we can determine the law of refraction for flux lines. It is evident from the figure that

$$\frac{\tan \alpha_1}{\tan \alpha_2} = \frac{\mu_1}{\mu_2}.$$

In passing from air into iron, the magnetic flux lines are deflected from the normal to a considerable extent and as a result the flux density sharply increases.

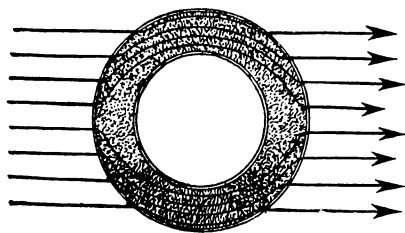


Fig. 120

For this reason, an iron body, whose permeability is hundreds or thousands of times greater than  $\mu_0$  "absorbs", flux lines. This is the basis of magnetic shielding. Magnetic flux cannot, in effect, penetrate into a region bounded by iron since practically all of the magnetic lines enter the iron (Fig. 120).

The distortion of a magnetic field due to a magnetic body of specified shape is determined in exactly the same manner as in the case of a dielectric. If the iron

body has the form of an ellipsoid, cylinder or plate, theoretical calculations show that the field inside the body will be uniform if the field was uniform before the iron was introduced. A relationship completely analogous to that given in Sec. 96 exists between  $H_0$ , the external uniform field before the body is introduced, and  $H_i$ , the field inside the iron after the body is introduced. The field intensity inside the iron body becomes less than the original intensity by an amount proportional to the magnetisation:

$$H_i = H_0 - \frac{N'}{\mu_0} J.$$

In order for the demagnetisation factor to be dimensionless, it is necessary to divide the magnetisation by the magnetic permeability of vacuum. Continuing with the SI system and substituting

$$J = \mu_0 (\mu - 1) H_i,$$

we obtain the following relationship between the external and internal fields:

$$H_i = \frac{H_0}{1 + (\mu - 1) N'}.$$

In the CGS system of units

$$H_i = H_0 - NJ,$$

$$J = \frac{\mu - 1}{4\pi} H_i,$$

and the relationship between the external and internal fields is given by

$$H_i = \frac{H_0}{1 + (\mu - 1) \frac{N}{4\pi}}.$$

The demagnetisation coefficient has the same value as in the case of dielectrics:

$$N = \frac{4\pi}{3} \left( N' = \frac{1}{3} \right)$$

for a sphere,  $N = 4\pi$  ( $N' = 1$ ) for a plate, etc.

#### Sec. 110. MAGNETIC HYSTERESIS

In discussing the permeability of iron, we may have created the false impression that the magnetic properties of ferromagnetic substances differ from those of paramagnetic substances only as regards the magnitude of the permeability. This is by no means the case. Ferromagnetic bodies differ from the others mainly in that the magnetic state of such a body is not linearly dependent, and moreover is not uniquely dependent, on the magnetic field intensity. Therefore, the concept of permeability for ferromagnetic substances is very relative. The magnetic properties of iron are best illustrated by a magnetisation vs. field intensity curve or a magnetic induction vs. field intensity curve. These curves are very similar to each other.

Let us consider the magnetisation of an iron body as a function of the field intensity. At first, the magnetisation increases slowly, then rapidly, and finally magnetic saturation sets in. Such magnetisation curves were first used by A. G. Stolotov and are characteristic of all ferromagnetic bodies (Fig. 121). We reiterate: the magnetisation and magnetic induction curves are very similar. The slope of the magnetisation curve gives the magnetic susceptibility, while the slope of the induction curve gives the magnetic permeability. From the figure, it is seen that the permeability (also susceptibility) curve has a maximum. For weak fields the permeability is low. As the field intensity is increased,  $\mu$  increases to a maximum, then begins to drop, and after reaching saturation remains unchanged. When the value of the permeability is given without specifying the external conditions, the maximum permeability is usually meant.

But there is more to be said about the behaviour of ferromagnetic substances. Let us assume that after the iron has been brought to a state of magnetic saturation the magnetic field intensity is decreased. It turns out that the induction decreases along a different curve, i.e., a curve lying higher than the initial magnetisation

curve. The field intensity may be decreased to zero, but magnetisation remains. The corresponding values of magnetisation and induction at this point are referred to as the *residual* values. To remove the residual magnetisation, the polarity of the field must be changed. Thus, in the case of the experiment discussed on p. 215, it would be necessary to reverse the polarity of the current in the primary winding. Demagnetisation occurs when the field intensity attains a value  $H_c$ , the value of the so-called *coercive* force. As the current is increased further in the same direction, the body begins to become magnetised in the reverse direction, i.e., what was previously a south pole becomes a north pole. The magnetic flux increases until the magnitude of the saturation is the same as in the initial process.

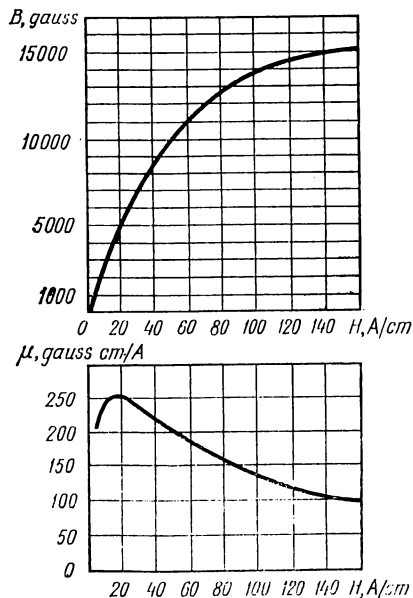


Fig. 121

400 Oe—one occurs during initial magnetisation, the second during demagnetisation and the third when the magnetisation process is repeated, i.e., just before

Having attained a negative induction maximum, we may then proceed with the process in the other direction. In this manner, the hysteresis loop shown in Fig. 122 is obtained.

It is seen from the figure that knowledge of the intensity of the field in which the iron is located is not sufficient to determine the magnetic induction and, hence, the magnetic permeability. Thus, for example, it is seen that three values of induction are possible for  $H =$

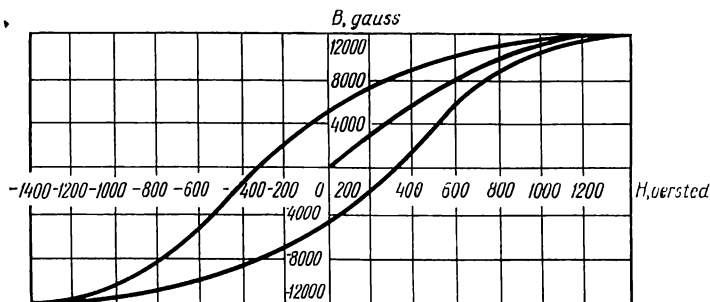


Fig. 122

the hysteresis loop is closed. The value of the magnetic induction, and the magnetic permeability, depends on the previous "history" of the sample. Hence, the designation "hysteresis loop".

A hysteresis loop is usually drawn on the assumption that the ferromagnetic body is brought to magnetic saturation. However, we can clearly obtain numerous hysteresis loops having smaller dimensions by inscribing them, as it were, in the fundamental loop. For this purpose, it is necessary to begin the demagnetisation



process before reaching saturation. Then, to each  $H$  value there corresponds an infinitely large number of values  $B$ .

A procedure based on this fact is used to bring a ferromagnetic body to a state in which both induction and field intensity are equal to zero. This "zero point" is achieved by a series of successive magnetic reversals, whereby each succeeding cycle is begun at a lower intensity level than the previous one.

The magnetic state of iron cannot be characterised only by the value of the permeability, field intensity or induction. Two of these quantities must be known—e.g., the induction and the intensity. The magnetic state of the iron is then represented by a point inside the fundamental hysteresis loop.

The nature of a hysteresis loop depends to a great extent on the material. A body is said to be magnetically soft if the coercive force (and hence the loop area) is small. Typical soft materials are pure iron, silicon steel, and iron and nickel alloys (particularly permalloy—78% nickel). Carbon and other steels belong to the magnetically hard materials and are used in the manufacture of permanent magnets.

Experiments show that the temperature of a ferromagnetic substance rises when it is subjected to magnetic reversals. This is very important in electrical engineering, for when iron is placed in a variable magnetic field the point on the  $B = f(H)$  curve representing the magnetic state of the iron is continuously tracing a hysteresis loop. Every time a loop is traced a certain amount of heat is released, which according to magnetic field theory is related to the loop area; of course, the lower the value of the induction maximum, the smaller the loop area. Therefore, empirical formulas may be sought relating heat released and maximum induction. In electrical engineering, for example, the following formula is widely used:

$$Q = \eta B_{\max}^{1.6},$$

where  $\eta$  is a coefficient whose value is given in tables.

*Example.* For a good transformer steel,  $\eta = 0.0011$ . When  $B_{\max} = 10,000$  G, the losses are equal to  $Q = \eta B_{\max}^{1.6} = 2.5 \times 10^3$  ergs/cm<sup>3</sup> =  $2.5 \times 10^{-4}$  J/cm<sup>3</sup>. This means that for magnetic reversals in iron due to an alternating current whose frequency  $\nu$  is 50 Hz the power loss is equal to  $12.5 \times 10^{-3}$  W for cubic centimetre of iron.

# Electromagnetic Fields. Maxwell's Equations

## Sec. 111. GENERALISATION OF THE LAW OF ELECTROMAGNETIC INDUCTION

In the preceding chapter it was shown that the motion of a conductor in a magnetic field is accompanied by induction phenomena. If this moving conductor is part of a circuit through which the magnetic flux changes when the conductor moves, a current corresponding to the induced emf  $\mathcal{E} = -\frac{1}{c} \frac{d\Phi}{dt}$  flows in the circuit.

This current is due to the action of a Lorentz force: a force equal to  $\frac{1}{c}[\mathbf{vB}]$  acts on a unit electric charge (the CGS system).

The induced current depends only on the relative displacement of the conductor with respect to the magnetic field. Thus, it may be asserted with equal validity that a Lorentz force is produced when a charge moves in a magnetic field or when the charge is "at rest" and the magnetic field moves. This follows from the principle of relativity.

Consider a system of coordinates relative to which a magnetic field moves. Such a coordinate system may be fixed, for example, relative to a laboratory bench along which the pole of a permanent magnet moves. Then, a Lorentz force will act on charges at rest relative to this bench. Let us assume nothing is known about the moving permanent magnet. Having established that a force acts on the stationary electric charges, we are perfectly justified in concluding that an electric field exists in this system whose intensity is equal to the Lorentz force divided by the magnitude of the charge. Thus the electric field intensity in the "stationary" coordinate system, relative to which the source of constant magnetic field moves with velocity  $\mathbf{v}$ , is expressed by the formula

$$\mathbf{E} = \frac{1}{c} [\mathbf{vB}].$$

Of course, the relations differ for the electric field created by charges and the electric field created by the motion of the system relative to the magnetic field. To begin with, this new field that we are considering has no charge sources. This means that the flux lines have neither beginning nor end. Moreover, it is not difficult to see that the flux lines of this electric field form closed curves, i.e., the electric field created by the moving magnetic field is rotational.

Imagine an arbitrary circuit that is stationary relative to the laboratory bench. The moving magnetic field crosses this circuit. If this imaginary circuit is replaced by a real wire circuit, then in accordance with Faraday's law an emf is induced in the circuit that is equal to  $\oint \mathcal{E} d\mathbf{l}$ . Hence, the integral  $U = \oint \mathbf{E} d\mathbf{l}$  is not equal to zero. This means that the electric field  $\mathbf{E} = \frac{1}{c} [\mathbf{vB}]$  created by the moving magnetic field is a rotational field.

For a real wire circuit  $U = \frac{1}{c} \frac{d\Phi}{dt}$ , where  $\Phi$  is the magnetic flux passing through the circuit. However, it is immaterial whether or not wire is present at the location of the closed curve. The equation  $U = \frac{1}{c} \frac{d\Phi}{dt}$  is also valid for an imaginary circuit in the region where the sources of the magnetic field are moving.

One final generalisation remains to be made. Experiments show that the cause for the change in the magnetic field is of no importance in the induction effect. The field change due to the motion of a permanent magnet and that due to a change of current strength in a stationary coil can always be made equal by, for example, bringing the permanent magnet closer or increasing the current in the coil creating the field. Therefore, the law under consideration must be valid in all cases, no matter how the magnetic field is changed. Thus, if the magnetic field (magnetic flux) changes in a certain region in space, a rotational electric field is produced that is related to the magnetic field change as indicated by the following law. The electromotive force  $U = \oint \mathbf{E} d\mathbf{l}$  along a closed curve is equal to the derivative with respect to time of the magnetic flux passing through this circuit:

$$U = \frac{1}{c} \frac{d\Phi}{dt}$$

or, in the SI system,

$$U = \frac{d\Phi}{dt}.$$

This is the *generalised law of induction*, one of the most important laws of nature. Let us examine the mathematical expression for this law. Equating the expressions for the electromotive force and the magnetic flux, the law can be written as follows:

$$\oint \mathbf{E} d\mathbf{l} = -\frac{1}{c} \frac{d}{dt} \int B \cos \alpha dS \quad (\text{CGS}),$$

$$\oint \mathbf{E} d\mathbf{l} = -\frac{d}{dt} \int B \cos \alpha dS \quad (\text{SI}).$$

First, as regards the minus sign that appears in this equation, it should be noted that in vector analysis the direction in which the circuit is traversed and

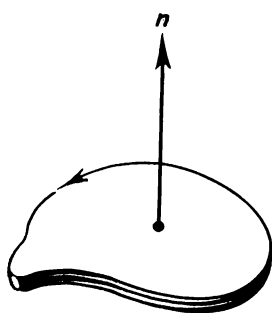


Fig. 123

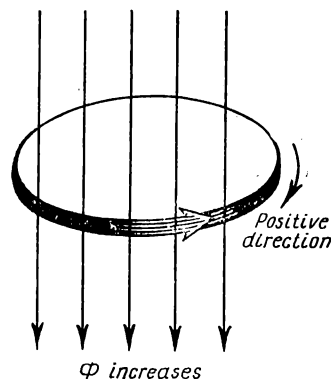
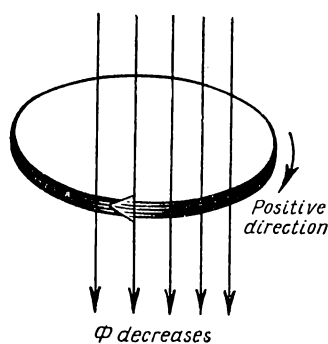


Fig. 124

the direction of the normal to the plane of the circuit are related to each other as follows: the positive direction of the normal in a right-handed screw system is such that as viewed from the vector terminus the circuit appears to be traversed in the counterclockwise direction (Fig. 123). Let us construct a closed curve in space and ascribe an arbitrary direction to it. The direction of the normal to the area encompassed by the curve under consideration is thus determined. Magnetic flux

passes through this circuit. At a given instant, it may be positive or negative, depending on whether the induction vector forms an acute or obtuse angle with the normal. The derivative of the flux with respect to time is positive if the flux is increasing and is negative if the flux is decreasing. Thus, taking the minus sign into account in the induction formula, the law may be stated as follows: the electromotive force is positive if positive flux is decreasing or negative flux is increasing, i.e., the direction of the electric lines of force coincides with the adopted direction for positive. On the other hand, the electromotive force is negative if positive flux is increasing or negative flux is decreasing. These relationships are clearly illustrated in Fig. 124.

We shall show that the minus sign in the induction formula is the mathematical expression of Lenz's law. Assume, for example, that the north pole of a bar magnet

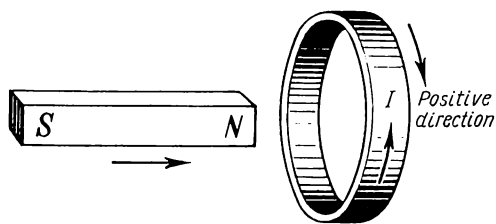


Fig. 125

approaches a coil and that the positive direction in the circuit is as indicated in Fig. 125. Then, the magnetic flux is positive and so is its derivative with respect to time. The electromotive force must be negative and the induced current is opposite to the direction adopted as positive. We can immediately find the magnetic field of the induced current by recalling that the flux lines emerge from the side of the current ring from which the current appears to be moving in a counterclockwise direction.

Therefore, as the magnet approaches the circuit, a current of such direction is induced in the latter that the field produced tends to oppose the action which caused it. This is a statement of Lenz's law. It is not difficult to verify this important rule for other particular cases as well.

Let us sum up. A varying magnetic field is inseparable from an electric field. Moreover, it is seen that the division of fields into electric and magnetic is a relative matter. From one viewpoint there is only a magnetic field in space. From another viewpoint in addition to the magnetic field there is an electric field.

A rotational electric field consists of electric lines of force that link the magnetic induction vectors when the magnetic flux passing through a closed line of force changes with time. When the flux increases, the direction of the line of force is clockwise as viewed from the induction vector terminus.

## Sec. 112. DISPLACEMENT CURRENT

Electromagnetic field theory, whose foundation was laid by Faraday, was mathematically perfected by the English scientist James Clerk Maxwell. One of Maxwell's most important new ideas was that symmetry must exist in the interdependence between magnetic and electric fields.

In the preceding article, we discussed the problem of creating an electric field by varying magnetic flux. The question naturally arises: does a variable flux of electric lines of force create its own magnetic field? Maxwell answered this question in the affirmative and advanced the hypothesis that a relationship exists between variable electric flux and a magnetic field that is quite analogous to the generalised law of induction. According to the hypothesis, if a change in electric flux occurs in some region of space, a rotational magnetic field is created. Moreover, the magnetomotive force  $U$  taken along a closed curve is equal to the change in elec-

total flux passing through this closed curve, i.e.,

$$U = \frac{dN}{dt},$$

where

$$U = \oint \mathbf{H} \, d\mathbf{l}$$

and the electric flux

$$N = \int_S \mathbf{D} \cos \alpha \, dS.$$

In the CGS system

$$U = \frac{1}{c} \frac{dN}{dt}.$$

The parallel in the relationships between magnetic and electric fields does not extend to the sign before the derivative of the flux.

As is known, when currents are present the magnetomotive force along a closed curve is equal to  $U = I$  (or  $\frac{4\pi}{c}I$  in the CGS system). How should the equation for magnetomotive force be written for such a closed curve that encloses electric current and variable flux of electric lines of force? Maxwell assumed that the magnetomotive forces are additive. Thus, the general formula has the form

$$\oint \mathbf{H} \, d\mathbf{l} = I + \frac{dN}{dt} \quad (\text{SI})$$

or

$$\oint \mathbf{H} \, d\mathbf{l} = \frac{4\pi}{c} \left( I + \frac{dN}{dt} \right) \quad (\text{CGS}).$$

The expression  $\frac{dN}{dt}$  has the dimensionality of electric current strength. Maxwell called it *displacement current*. He thereby incorporated in this designation the very widespread notion at the end of the nineteenth century that the field in a vacuum displaces the particles of an "ether" from their positions of equilibrium. This designation has continued to prevail in science, although now the presence of a field in vacuum is not related to the concept of particle displacement of any medium whatsoever. In a dielectric medium, the displacement current  $\frac{dN}{dt}$  may be resolved into two components, corresponding to the intensity and polarisation vectors into which the displacement vector  $\mathbf{D}$  can be resolved (see p. 193). Therefore, the portion of the displacement current "flowing" in the dielectric is determined by the change in the polarisation vector, i.e., by the relative displacements of the centres of gravity of the positive and negative charges.

Before discussing the role of displacement current in one or another process, we shall prove an important proposition concerning the sum of the conduction and displacement currents.

Consider an arbitrary system of electric currents and imagine a closed surface drawn in such a manner that the currents intersect it. If the currents are constant it follows directly from the law of conservation of electricity that the sum of the currents entering the closed surface must be equal to the sum of the emerging currents, or, to be more concise, the algebraic sum of the currents flowing through a closed surface is equal to zero. It is evident that this law may not be obeyed by variable currents: for example, in the case of a closed surface enveloping one plate

of a condenser connected in an alternating current circuit (Fig. 126) or a closed surface through which the top of an antenna protrudes at one point.

However, this theorem is valid for variable currents as well if the term "current" is taken to mean "total current", i.e., the conduction current plus the displacement current, rather than the conduction current alone. To prove this, it suffices to consider an arbitrary curve on which a surface is based and for which the following relation is valid:

$$\oint \mathbf{H} d\mathbf{l} = I + I_{\text{displ.}}$$

By gradually reducing the closed curve to zero, the surface  $S$  on which this circuit is based becomes closed as shown in Fig. 127. (This process is similar to con-

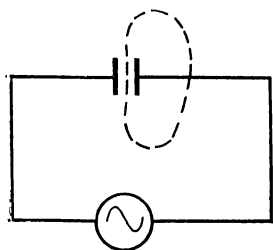


Fig. 126

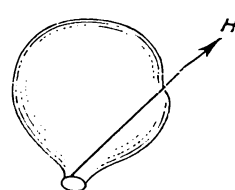
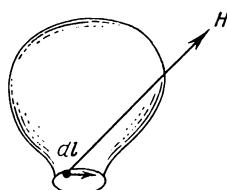
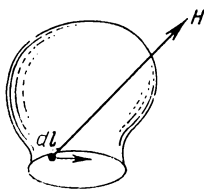


Fig. 127

stricting the opening of a draw-string pouch.) The magnetomotive force reduces to zero and hence the sum of the conduction and displacement currents passing through the closed surface also equals zero.

Now let us discuss the role of displacement currents in electromagnetic phenomena

It can be shown that displacement currents are negligibly small where the conduction currents are different from zero. Hence, the displacement current within a conductor is always disregarded.

In calculating the value of the displacement current in dielectrics, two cases should be considered—displacement currents in a dielectric surrounded by a conductor forming a closed circuit and displacement currents that are a continuation of a conductor of an open circuit.

Consider a conductor forming a closed circuit in which electric current is flowing and which is intersected by a closed surface. If the current is constant, then at each instant the same amount of electricity passes outward through the surface as inward. The situation is different in the case of variable currents. Here, the strength of the variable current may have different values in different parts of the circuit (see below, p. 244). As a result, the strengths of the currents passing inward and outward through the surface at a given instant may not be equal. Then, a displacement current "flows" via the dielectric from the point where the current is less to the point where the current is greater, and in this manner the current deficit is compensated for. Clearly, changes in the displacement current with respect to time will exactly correspond to changes in the conduction current. This phenomenon is significant only when the frequency of the current is sufficiently high.

If the conduction current does not form a closed circuit, e.g., in the case of an alternating current circuit containing a condenser, the conduction and displace-

ment currents are simply equal to each other. In this case, it may be said that the conduction current circuit is closed by the displacement current.

In spite of the fact that the magnitudes of the displacement currents are quite large for such cases, certain calculations may be performed without considering them. Thus, when the conduction current circuit is closed by the displacement current between the condenser plates, the magnetic field created in this region by the displacement current is the same as the field that would have been produced if the conduction current flowed in an uninterrupted circuit. Therefore, the presence of displacement current does not affect the calculation of the magnetic field, the coefficient of self-induction of a system, etc.

#### Sec. 113. NATURE OF AN ELECTROMAGNETIC FIELD

The following equations, which were discussed in the two preceding articles, are called *Maxwell's equations*:

$$\oint \mathbf{E} d\mathbf{l} = -\frac{d\Phi}{dt} \quad \text{and} \quad \oint \mathbf{H} d\mathbf{l} = I + \frac{dN}{dt}.$$

These equations concisely sum up our knowledge of electromagnetic fields.

Maxwell's equations cannot be derived. The discussion of the two preceding articles does not constitute a derivation, but constitutes rather an illustration of conjectures leading Maxwell to his discovery.

A large class of phenomena of interest to physicists, electrical engineers and radio engineers obey Maxwell's equations. The laws of these phenomena are a consequence of Maxwell's equations and may be derived from them. The extraordinary importance of the predictions based on Maxwell's equations gives these equations equal rank with Newton's laws of motion and the principles of thermodynamics as fundamental laws of nature.

We shall not go into the mathematical methods of solving Maxwell's equations. It turns out that the above integral equations may be transformed into differential equations. Then, by solving Maxwell's differential equations, it is possible in principle to determine the electromagnetic field for a given distribution of charge and current.

Let us again consider the physical essence of electromagnetic phenomena as given by Maxwell's equations. It may be summarised in the following manner.

The division of an electromagnetic field into electric and magnetic fields has only relative meaning. If from the viewpoint of an inertial system of coordinates only a magnetic field exists, then from the viewpoint of another system moving relative to this system there exists an electric field in addition to a magnetic field. The converse is also true, namely, if an observer in one system of coordinates finds only an electric field present, then an observer in another inertial system will find that both an electric field and magnetic field exist.

Let us now consider an electromagnetic field from the viewpoint of an inertial frame of reference, first directing our attention to a region of space in which free electric charges and hence conduction currents are absent. In this case, Maxwell's equations have the form

$$\oint \mathbf{E} d\mathbf{l} = -\frac{d\Phi}{dt} \quad \text{and} \quad \oint \mathbf{H} d\mathbf{l} = \frac{dN}{dt}.$$

Both the magnetic and electric fields have a purely rotational character, i.e., the lines of force are closed and mutually interwoven: electric lines encircle magnetic lines and magnetic lines encircle electric lines. The electromagnetic field may be

depicted as a chain of rings, whereby closed magnetic lines of force are alternately linked with closed electric lines of force (Fig. 128). Such a chain exists only if the field is variable: a ring of increasing magnetic flux creates about itself a ring electric flux; the varying electric field creates a ring of magnetic flux, etc.

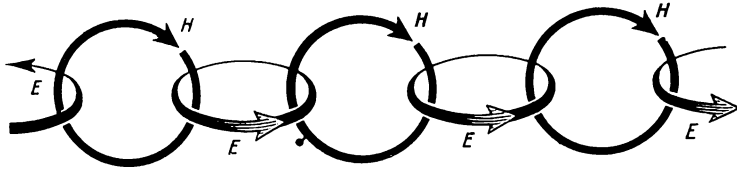


Fig. 128

If the region of space under consideration contains charges and currents, there in addition to rotational fields with linked lines of force there exists a rotational magnetic field whose closed flux lines encircle currents and a potential electric field whose flux lines begin on positive charges and terminate on negative charges.



## Energy Transformations in Electromagnetic Fields

### Sec. 114. TRANSFORMATIONS IN STEADY CURRENT CIRCUITS

Let us consider a portion of a conductor through which a steady electric current is flowing. If the resistance of the conductor segment is  $R$  and the potential difference across it is  $U$ , the current strength is determined by Ohm's law, namely,  $I = \frac{U}{R}$ . The electric field performs work in moving charges along the circuit and the work per unit charge is equal to  $U$ . Since the current strength is, by definition, the quantity of electric charge flowing through the conductor cross-section per unit time, the product  $IU$  yields the work performed by the field in moving the electric charge per unit time. This product,  $IU$ , represents power. If the current is a steady one, this work is completely converted into heat<sup>†</sup> (Joule heat). Thus, the formula for calculating the thermal effect of current is

$$IU = \frac{U^2}{R} = I^2 R.$$

The transformation of the work performed by the electric field into heat occurs at each point of the conductor. To express this mathematically, Ohm's law must be converted into a form applicable to a point of a conductor, rather than to a portion of a conductor. By introducing the current density  $j$ , which is equal to  $\frac{I}{S}$  where  $S$  is the cross-section of the conductor, and by replacing the expression for the potential difference by  $El$ , and, finally, by expressing the resistance in terms of the conductor length  $l$  and its cross-section, i.e.,  $R = \frac{1}{\lambda} \frac{l}{S}$ , we obtain:  $j = \lambda E$ .

Thus, it may be said that the current density is directly proportional to the electric field intensity. The specific conductivity  $\lambda$  is the coefficient of proportionality and the direction of the current is assumed to coincide at each point with the direction<sup>‡</sup> of the intensity. The formula

$$\mathbf{j} = \lambda \mathbf{E}$$

is called *the differential form of Ohm's law* and should be viewed as an empirical law generalising the laws for current flow in conductors. Ohm's law in its usual (integral) form is a consequence of this equation.

Consider an infinitely small volume element of the conductor,  $d\tau$ , in the form of a cylinder whose generatrix  $dl$  is parallel to the flux lines and whose base  $dS$  is perpendicular to the current. The amount of electric charge flowing through a cross-section of the cylinder is  $j dS$  and the potential difference between the ends of the element is  $E dl$ . Hence, the work performed by the field in moving the electric charge through this volume is equal to  $j E d\tau$ . This formula also gives the heat released inside the volume  $d\tau$ . If we are interested in the work of the current in a small volume of the conductor, the last expression must be integrated. The formula

$$j E = \frac{j^2}{\lambda} = \lambda E^2$$

gives us the expression for the work of the current or the Joule heat released per unit volume of the conductor.

Thus, in the case of a portion of a direct-current circuit, the energy transformations are reduced to the transformation into heat of the work done by a field. However, the picture changes as regards the energy balance of the entire closed direct-current circuit. The work performed by the electric forces along a closed curve when the field is constant is equal to zero, for the work performed by the electric forces in moving charge along the external portion of the circuit is equal and opposite to the work required to move the charge along the internal portion of the circuit. Therefore, the release of Joule heat in a direct-current circuit occurs only at the expense of the energy supplied by the current source—accumulator, electric generator, etc.—i.e., at the expense of energy of nonelectrical origin or, as it is sometimes said, at the expense of the energy of an “applied” force. The role of electric current is reduced simply to the “transfer” of energy from the current source to the point where the heat is released. The energy the source is able to supply is given by the electromotive force  $\mathcal{E}$ , which by definition is measured by the work performed in moving a unit charge along a closed curve. Actually, the applied emf performs this work only over those small portions of the circuit where the charge must be moved against the forces of an electric field.

The power of the direct-current circuit is given by the expression  $I\mathcal{E}$ . This may be expressed in terms of unit volume if it is assumed that the applied forces are distributed throughout the volume. Then, the work performed by the applied forces is given by

$$jE^{appl}$$

where  $E^{appl}$  is the “intensity” of the applied forces.

Designating the work of the applied forces by  $P$ , and the Joule heat released by  $Q$ , the essence of the electrical transformations in a direct-current circuit may be expressed by the concise formula

$$P - Q = 0.$$

*Example.* For an isolated copper wire of cross-section  $S = 4 \text{ mm}^2$ , the permissible current density in the case of an open wire is  $j = 9 \times 10^6 \text{ A/m}^2$ . A 1-metre length of such a wire has a resistance of  $4.25 \times 10^{-3} \text{ ohm}$ . For the indicated value of  $j$ , a current of  $I = 36 \text{ A}$  flows in the wire and the Joule heat losses per second for this portion of the circuit amount to  $I^2 R = 1,296 \times 4.25 \times 10^{-3} \approx 6 \text{ J}$ , i.e., in a unit volume  $0.33 \text{ cal} = 1.38 \text{ J}$  are released each second.

#### Sec. 115. TRANSFORMATIONS IN A CLOSED CIRCUIT OF VARIABLE CURRENT

A flow of variable current is inevitably accompanied by induction effects. Thus, to a variable current strength there corresponds a variable magnetic flux  $\Phi$ . Here,  $\Phi$  represents the number of lines of force which are created by the current circuit and which pass through the conducting circuit. In this case, the induction effects are due to the current's own magnetic flux, whence the designation *self-induction*. Since  $\Phi$  is continuously changing, an induced emf  $\mathcal{E}^{\text{ind}} = -\frac{d\Phi}{dt}$ , exists in the current circuit at each instant, in addition to the applied emf.

The magnetic flux is always proportional to the first power of the current. Hence, the formula  $\Phi = LI$  has universal validity. The coefficient  $L$  is called the *inductance* of the circuit or the *coefficient of self-induction*. The value of  $L$  depends on the geometric properties of the circuit and the nature and distribution of magnetic bodies in the system. It does not depend on the conditions under which the system of conductors and magnetic bodies operate. Thus, for self-induced emf

the following equation holds:

$$\mathcal{E}^{ind} = -L \frac{dI}{dt}.$$

The significance of the minus sign in this formula may be explained as follows: when the current increases, the induced emf opposes the applied force, i.e., the induction is in the direction opposite to the applied force. On the other hand, when the current decreases, the directions of the induced emf and the applied emf coincide. This is the reason for the analogy generally made between mechanical inertia and self-induction. Self-induction impedes an increase as well as a decrease in current.

Ohm's law, relating emf and current strength, remains valid. Therefore, the product of current strength and total circuit resistance will at each instant be given by the following relation:

$$IR = \mathcal{E}^{appl} + \mathcal{E}^{ind} = \mathcal{E}^{appl} - L \frac{dI}{dt}.$$

Multiplying both members of the equation by the instantaneous current strength, we obtain the energy equation:

$$I^2R = I\mathcal{E}^{appl} - LI \frac{dI}{dt}.$$

Here,  $I\mathcal{E}^{appl} = P$  is the work of the applied forces and  $I^2R = Q$  is the Joule heat. It is seen that in a variable current circuit these two quantities are not equal to each other. The difference  $P - Q$  is equal at each instant to  $LI \frac{dI}{dt}$ , i.e., it is equal to the derivative of  $\frac{1}{2} LI^2$ . In other words, the excess of the work of the applied forces over the Joule heat released goes to increase the magnitude of  $\frac{1}{2} LI^2$ . On the other hand, the excess of the heat released over the work of the applied forces occurs at the expense of the magnitude of  $\frac{1}{2} LI^2$ . The equation

$$P - Q = \frac{d}{dt} \left( \frac{1}{2} LI^2 \right)$$

is the expression for the law of conservation of energy.

Clearly, the quantity  $W = \frac{1}{2} LI^2$  represents energy. It is the magnetic energy of a system that is inseparably linked with the existence of a magnetic field in it. (In the CGS system the expression for magnetic energy is  $\frac{1}{c^2} \frac{1}{2} LI^2$ ). There is magnetic energy in a direct-current circuit too, but in this case it does not manifest itself since it remains unchanged. The induction effects occur only when the current is switched on and off. When the circuit is closed the applied forces perform work, which is expended not only in the release of heat, but also in the storage of magnetic energy. On the other hand, when the circuit is opened the thermal energy released is at the expense of the magnetic energy of the current.

The magnetic energy formula may be verified experimentally by closing or, even better, opening a current circuit. The thermal energy released after the source is disconnected is numerically equal to the magnetic energy of the current. If the coefficient of self-induction is large, the release of heat continues over a period of time sufficiently long to enable us to measure the heat by, for example, calorimetric means.

Inductance may be measured in various ways and in the simplest cases may be calculated from the formula  $L = \frac{\Phi}{I}$ . The problem is reduced to the calculation of the magnetic flux passing through the system.

An expression for the inductance of a ring solenoid will be required below. The magnetic flux through one turn of a coil is  $\Phi = \mu_0 \mu H S$ , where  $S$  is the area of the turn, and the flux through  $n$  turns is  $\Phi = n \mu_0 \mu H S$ . Substituting the expression for the field intensity (using the practical system of units), we obtain

$$\Phi = n \mu_0 \mu S \frac{nI}{l}$$

Now, dividing both members of this equation by the current strength, we obtain an expression for the inductance of a coil (also approximately valid for an open solenoid):

$$L = \mu_0 \mu \frac{n^2}{l} S.$$

The inductance of a coil is directly proportional to the magnetic permeability of the medium and increases sharply with the number of turns. An increase in inductance is achieved by using iron or by increasing the number of turns. In order to make clear the relationship existing between the coefficient of self-induction and the dimensions of the coil, let us multiply the numerator and the denominator by  $l$ . Then,

$$L = \mu_0 \mu \left( \frac{n}{l} \right)^2 V,$$

and it is evident that the inductance is directly proportional to the volume occupied by the magnetic field and to the turn "density" squared.

*Example.* Consider a long solenoid of small cross-section ( $l = 15$  cm,  $n = 1,500$  turns,  $S = 1$  cm<sup>2</sup> and  $I = 0.1$  A). At the centre of the solenoid, the magnetic flux is  $\Phi = n \mu_0 \mu H S = 6\pi \times 10^{-6}$  V s. The inductance of this solenoid is

$$L = \mu_0 \mu \frac{n^2}{l} S = 4\pi \times 10^{-7} \times 1 \times \frac{(1,500)^2}{0.15} \times 10^{-4} = 1.9 \times 10^{-3} \text{ henry (H)}.$$

In the SI system the inductance is measured in henrys (1 henry = 1 ohm s). The inductance of coils in radio engineering is measured in millionths and thousandths of a henry. Chokes having iron cores can attain inductance values of the order of a number of henrys.

#### Sec. 116. MAGNETIC ENERGY OF A FIELD

In the chapter devoted to electric fields, it was shown that the electric energy of a system may be viewed as a quantity whose density distribution is represented by  $\frac{1}{2} \epsilon E^2$  (in the SI system). The electric energy of the system is then determined by integrating this expression over the region occupied by the field. The importance of this circumstance was emphasised as it enables us to express the energy in terms of the field intensity and it confirms our conception of a field as something that can be localised.

Naturally, we expect the situation to be similar for magnetic fields and this is indeed the case. It can be mathematically shown that the transition from the magnetic energy formula,  $\frac{1}{2} LI^2$ , to the expression for the magnetic energy density,  $\frac{1}{2} \mu_0 \mu H^2$ , is completely analogous to the corresponding transition from electric fields.

Let us consider this transition for the simple case of the uniform field of a ring solenoid. Substituting the expression for the inductance in the magnetic energy formula, we obtain

$$W_M = \frac{\mu_0 \mu \left( \frac{n}{l} \right)^2 I^2}{2} V$$

But  $\frac{nI}{l}$  is the field intensity. Hence, the magnetic energy of a coil may be written in the form

$$W_M = \frac{\mu_0 \mu H^2}{2} V,$$

so that the magnetic energy density is given by the expression

$$w_M = \frac{\mu_0 \mu H^2}{2} \quad (\text{SI}),$$

$$w_M = \frac{1}{8\pi} \mu H^2 \quad (\text{CGS}).$$

Thus, for any system of currents, the magnetic energy may be represented by the integral over the volume occupied by the field:

$$W_M = \frac{\mu_0}{2} \int \mu H^2 d\tau \quad (\text{SI})$$

and

$$W_M = \frac{1}{8\pi} \int \mu H^2 d\tau \quad (\text{CGS}).$$

Now, consider the magnetic energy of two currents. The expression for this energy naturally divides into three integrals if the intensity of the resultant field  $H$  is viewed as the sum of the field intensities of the two currents:  $H = H_1 + H_2$ . In the following expression for the magnetic energy, the significance of each of the integrals is quite evident:

$$W = \frac{\mu_0}{2} \int \mu H_1^2 d\tau + \mu_0 \int \mu H_1 H_2 d\tau + \frac{\mu_0}{2} \int \mu H_2^2 d\tau.$$

The first and third integrals yield the magnetic energy of the first and second currents, respectively, while the second integral represents the interaction energy of the two currents. This last integral may assume different values even though the magnitudes of the field intensities,  $H_1$  and  $H_2$ , do not change. Thus, if the mutual disposition of the two currents changes, the field vectors  $H_1$  and  $H_2$  are turned, generally speaking, relative to each other and the value of the interaction energy also changes.

Of course, the first and third integrals may be expressed in terms of the current strength and the inductance:  $\frac{1}{2} L_1 I_1^2$  and  $\frac{1}{2} L_2 I_2^2$ . As for the second integral, it is clear that its value is proportional to the product of the current strengths. Thus,

$$\int \mu H_1 H_2 d\tau = M I_1 I_2.$$

The coefficient of proportionality  $M$  is known as the *coefficient of mutual induction*. Just as in the case of inductance,  $M$  depends on the geometry of the system and the distribution of magnetic bodies.

Thus, it is evident that the change in the magnetic energy of a system of currents is related not only to the work of the applied forces and the Joule heat

released, but also to the work performed by the field in moving the conductors under the action of an Ampère force. Hence, the law of conservation of energy requires that the following equation be satisfied:

$$dW_M = -A - (Q - P) dt,$$

where  $A$  is the mechanical work. Thus, it may be stated that, in the general case, the magnetic energy expended is equal to the work of moving the conductors and to the excess of released Joule heat over the work of the applied forces.

The relations introduced in this article do not take into account one phenomenon—magnetic hysteresis. This problem will not be dealt with because of its specialised nature.

*Example.* The energy stored in the magnetic field of the coil described in the example on p. 315 is

$$W_M = \frac{LI^2}{2} = \frac{1.9 \times 10^{-3} \times (0.4)^2}{2} = 0.95 \times 10^{-5} \text{ J}.$$

The energy density is

$$w_M = \mu_0 \mu \frac{H^2}{2} = 4\pi \times 10^{-7} \times 1 \times \frac{(1,000)^2}{2} = 0.63 \text{ J/m}^3.$$

Naturally, the same result could be obtained by dividing the total energy of the magnetic field by the volume of the coil:

$$w_M = \frac{W_M}{Sl}.$$

## Sec. 117. ELECTRIC OSCILLATIONS

The processes of transforming electric energy into magnetic energy and vice versa are of fundamental importance in electrodynamics. A simple system in which such transformations occur is a charged electric condenser whose plates are connected at a certain instant to the ends of a coil (Fig. 129). When the condenser discharges, an electric current flows through the coil and creates a magnetic field around it. At each instant, the electric field of the condenser and the magnetic field of the coil are closely linked. The energy of this system at each instant is equal to the energy of the electric field, which is concentrated mainly between the condenser plates, and the energy of the magnetic field, which is concentrated inside the coil. As is well known, electric oscillations

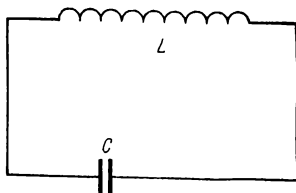


Fig. 129

arise in such a circuit and we shall now show that such oscillations are inevitable.

To begin with, let us disregard thermal energy losses. Then, the law of conservation of energy requires that the following equation be satisfied:

$$W = \frac{1}{2} \frac{Q^2}{C} + \frac{1}{2} LI^2 = \text{const.}$$

The sum of the electric energy and the magnetic energy is the same at each instant. Hence, the derivative of the above expression with respect to time is equal to zero:

$$\frac{dW}{dt} = \frac{Q}{C} \frac{dQ}{dt} + L \frac{I dI}{dt} = 0.$$

Since the current strength is equal to the decrease in charge on the condenser plates, i.e.,

$$I = \frac{dQ}{dt},$$

the equation may be simplified as follows:

$$\frac{Q}{C} + L \frac{dI}{dt} = 0.$$

Such a relationship between the charge on the plates of a condenser and the current strength, which is equal to the derivative of the charge with respect to time, can be satisfied only if harmonic oscillation of the charge and the current is assumed.

This becomes evident upon comparing the above relations with the equations for mechanical vibrations (see p. 65):

$$\begin{aligned} I &= \frac{dQ}{dt} & v &= \frac{dx}{dt}; \\ L \frac{dI}{dt} &= -\frac{1}{C} Q, & m \frac{dv}{dt} &= -kx. \end{aligned}$$

Charge and current, on the one hand, are analogous to displacement from equilibrium and velocity of motion, on the other. As for the parameters of the system—inductance is analogous to mass and reciprocal capacitance is analogous to the rigidity of the system.

Let the initial time equal the instant when the condenser is fully charged, and assume that

$$Q = Q_0 \cos \omega t.$$

Then,

$$I = -Q_0 \omega \sin \omega t.$$

Substituting in the differential equation, we obtain

$$-LQ_0 \omega^2 \cos \omega t = -\frac{1}{C} Q_0 \cos \omega t$$

or, after cancelling,

$$\omega = \frac{1}{\sqrt{LC}}.$$

Thus, irrespective of the initial charge on the condenser plates, the harmonic oscillations occurring in the condenser have a natural frequency  $\omega_0 = \frac{1}{\sqrt{LC}}$ . The smaller the capacitance and inductance of the circuit, the higher the frequency of the electric oscillations.

What is the situation in a real circuit where the Joule heat losses cannot be neglected? Clearly, the total energy of the system in this case will decrease in accordance with the equation

$$dW = -I^2 R dt, \quad \text{i.e.,} \quad -I^2 R = \frac{1}{C} Q \frac{dQ}{dt} + LI \frac{dI}{dt}.$$

Differentiating again with respect to time and using the relationship between charge and current, we obtain an equation of the form

$$L \frac{d^2 I}{dt^2} + R \frac{dI}{dt} + \frac{1}{C} I = 0.$$

At this point, an analogy should be drawn between the corresponding electric and mechanical quantities. Comparing the last equation with the equation for mechanical vibrations with friction (p. 69), it is seen that the electric resistance is analogous to the coefficient  $\alpha$ , which is a measure of the mechanical resistance.

The solutions of such linear differential equations are considered in courses in advanced mathematics. We shall simply give the final result, which incidentally

is easily verified by substitution in the above equation:

$$I = I_0 e^{-\beta t} \cos \omega t.$$

The frequency of oscillation is given by

$$\omega = \sqrt{\omega_0^2 - \beta^2}.$$

Thus, the process is determined by two characteristics—the natural frequency of free undamped oscillations,  $\omega_0 = \frac{1}{\sqrt{LC}}$ , and the damping coefficient,  $\beta = \frac{R}{2L}$ .

It is seen, first, that light damping is achieved by decreasing the resistance relative to the inductance. (To be sure, this is not easily done, for if we increase the number of turns of the coil, both quantities increase simultaneously. However,  $L$  does increase at a faster rate.) Secondly, it will be noted that when

$$\omega_0^2 < \beta^2, \text{ i.e., } 4L < CR^2,$$

oscillations become impossible. The discharge of the condenser under such conditions leads to an aperiodic process analogous to the return swing of a pendulum displaced in a viscous medium from its equilibrium position.

*Example.* Assume that we have a variable condenser whose maximum capacitance  $C = 500$  pF. Calculate the corresponding inductances of the radio coils for the 1,500 metre and 15 metre wavelengths.

1. The frequency of electric oscillations corresponding to  $\lambda_1 = 1,500$  m is  $\nu_1 = 2 \times 10^5$  Hz = 200 kHz. Since

$$\omega = 2\pi\nu_1 = \frac{1}{\sqrt{L_1 C}}, \quad \text{then} \quad L_1 = \frac{1}{4\pi^2\nu_1^2 C} = 1.2 \times 10^{-3} \text{ H} = 1.2 \text{ mH}.$$

In order for the process in the circuit to be periodic, the resistance of the circuit must be less than

$$R_1 = 2 \sqrt{\frac{L_1}{C}} = 3,000 \Omega.$$

2.  $\lambda_2 = 15$  m,  $\nu_2 = 2 \times 10^7$  Hz = 20 MHz,  $L_2 = 0.12 \times 10^{-6}$  H = 0.12  $\mu$ H. In order for oscillations to be possible, the resistance of the circuit must be less than  $R_2 = 2 \sqrt{\frac{L_2}{C}} = 30 \Omega$ .

#### Sec. 118. ELECTROMAGNETIC ENERGY

In a system in which the oscillatory circuit consists of a condenser and a coil (particularly if the condenser is composed of large plates separated by a short distance and the coil has a large number of turns), the electric and magnetic fields are concentrated in their respective regions. Therefore, it is possible to consider the electric and magnetic energies as two related, but nevertheless distinct quantities. This division loses physical significance to a large extent when we consider rapidly varying fields, where large electric and magnetic fields exist in the same region.

Recalling what was said in Sec. 113 about the relative nature of the division of an electromagnetic field into electric and magnetic components, it should be understandable that it is necessary to introduce into the theory the concept of an electromagnetic energy that is formally equal to the sum of the electric and magnetic energy of the field. The density of electromagnetic energy in space is

$$w = \frac{1}{8\pi} (\epsilon E^2 + \mu H^2),$$



while the electromagnetic energy contained in the volume  $V$  is

$$W = \frac{1}{8\pi} \int_V (\epsilon E^2 + \mu H^2) dV.$$

In rapidly varying fields, the physical significance of the transformation of magnetic energy into electric energy, and vice versa, is lost. At the same time, any energy transformations occurring in an electromagnetic field must be taken into account in the energy balance by a single electromagnetic energy quantity.

If the above expression for electromagnetic energy is assumed to be valid, then, using the electromagnetic field equations of the preceding chapter, the following theorem for the decrease in electromagnetic energy within a certain volume of space can be rigorously proved:

$$-\frac{dW}{dt} = (P - Q) + \oint K \cos \alpha dS.$$

This theorem was proved in 1884 by Poynting. (It was proved in a more general form, i.e., not in connection with an electromagnetic field, by N. A. Umov in 1874.) The integral on the right is the flux of the vector  $K^*$ . Using calculations that we have been forced to omit due to their complexity, it can be shown that this vector is perpendicular to the plane passing through the field vectors  $E$  and  $H$  (Fig. 130), and is equal to  $K = \frac{c}{4\pi} [EH]$  in the CGS system and to  $K = [EH]$  in the SI system.

Since the values of the field intensities decrease quite rapidly with increasing distance from the field sources, the flux of the Poynting vector is reduced to zero when all of space is taken into account. In this case, the theorem states: the change in electromagnetic energy is equal to the excess of the work of the applied forces over the released heat.

However, of most interest is the application of the theorem to a finite volume, i.e., when the flux of the Poynting vector is not equal to zero. If the volume under consideration does not include currents, the equation assumes the form

$$-\frac{dW}{dt} = \oint K \cos \alpha dS.$$

The change in electromagnetic energy is equal to the flux of the Poynting vector through the surface bounding the volume under consideration.

The Poynting vector characterises the flux of the electromagnetic energy and the last equation expresses the following fundamental concept: a change in electromagnetic energy within some volume is accompanied by an outflow or inflow of an equivalent amount of energy.

In essence, Poynting's theorem is a direct consequence of the law of conservation of energy and the postulate stating that electromagnetic energy can be localised in space.

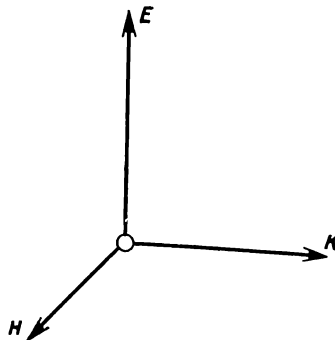


Fig. 130

---

\* It should be recalled that in mathematics an expression in the form  $\int_S A dS$  is called the flux of vector  $A$  through the surface  $S$ .

If Poynting's vector really has the significance of energy flux it should be related to the energy density as follows:  $K = v\omega$  (cf. p. 88, where an analogous problem is considered relative to the propagation of elastic waves in a medium). By means of Maxwell's theory, we can determine  $v$ , the propagation velocity of the electromagnetic energy. It turns out that

$$v = \frac{c}{\sqrt{\epsilon\mu}}.$$

Thus, in vacuum electromagnetic energy should be propagated at a velocity  $c = 3 \times 10^{10}$  cm/s, which agrees excellently with experiment. The coincidence between the values of  $c$  determined from purely electrodynamic experiments (e.g.

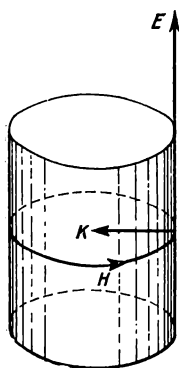


Fig. 131

the measurement of the interaction between two currents and the value of this constant determined by direct measurement of the propagation velocity of electromagnetic waves is remarkable and may be taken as practically conclusive proof of the validity of Maxwell's theory.

In a medium, the value of the propagation velocity of an electromagnetic wave is  $c$  divided by  $\sqrt{\epsilon\mu}$ . We shall see below under what conditions this relationship is satisfied and it will be explained why certain deviations occur.

Let us now return to the consideration of energy transformations in finite regions of space that contain conducting currents.

Assume that in the region under investigation there exists a cylindrical conductor of radius  $r$  through which a current of density  $j$  flows. The intensity of the magnetic field at the surface of the conductor (cf. p. 211) is equal in the CGS sys-

tem to  $H = \frac{2\pi}{c} rj$  and the magnetic flux lines form circles about the current axis. It can be seen (see Fig. 131) that Poynting's vector is directed into the conductor, for the field intensity and the current vector have the same direction. For the numerical value of Poynting's vector, we obtain (at the surface of the conductor):

$$\frac{c}{4\pi} EH = \frac{c}{4\pi} \frac{j}{\lambda} H = \frac{j^2 r}{2\lambda}.$$

Now, let us determine the flux of Poynting's vector in a conductor segment length  $l$ . This flux is equal to

$$K \times 2\pi r l = \frac{j^2}{\lambda} \pi r^2 l = \frac{j^2}{\lambda} V,$$

where  $V$  is the volume of the conductor segment. But  $\frac{j^2}{\lambda}$  is simply the thermal energy released per unit volume of conductor. We have thus shown that the flux of Poynting vector enters the conductor and transfers to it an amount of energy that is exactly equal to the energy expended in heat.

Where does this flux come from? In exactly the same manner as above, it can be shown that the flux of energy comes from those portions of the conductor where applied forces are present.

This picture explains how electromagnetic energy is propagated along conductors. When electric power is transmitted from Kuibyshev to a consumer in Moscow, the energy is delivered by electromagnetic waves and not by the first electrons initiating the motion along the conductor.

*Examples. 1.* Let us determine the order of magnitude of the electromotive force produced in the antenna of a radio receiver located at a distance  $R = 100$  km from a transmitter whose power is  $P = 100$  kW  $= 10^5$  J/sec.

The numerical value of Poynting's vector at the location of the receiving antenna is

$$K = \frac{P}{4\pi R^2} = \frac{10^5}{4\pi (10^5)^2} = 8 \times 10^{-7} \text{ J/m}^2 \text{ sec (W/m}^2\text{)}.$$

In the CGS system, the  $E$  and  $H$  vectors have the same dimensions ( $\text{g}^{1/2}\text{cm}^{-1/2}\text{sec}^{-1}$ ). It can be shown that for an electromagnetic wave propagating in vacuum the numerical values of the  $E$  and  $H$  vectors in the CGS system are equal:  $E = H$ . For these quantities, the following relationships exist in the SI and CGS systems:

$$1 \text{ V/m} = \frac{1}{3} \times 10^{-4} \text{ CGS unit/cm}; \quad 1 \text{ A/m} = 4\pi \times 10^{-3} \text{ Oe}.$$

Then the numerical values of the  $E$  and  $H$  vectors in the SI system are:

$$E' = \frac{1}{3} \times 10^{-4} E; \quad H' = 4\pi \times 10^{-3} H.$$

Therefore, for an electromagnetic wave ( $E' = H'$ ), we obtain:  $E = 120 \pi H$ . In the SI system  $K = EH$ ; hence,  $K = \frac{E^2}{120\pi}$  and  $E = \sqrt{120 \pi K} = 1.7 \times 10^{-2} \text{ V/m}$ .

This means that the potential difference produced in a receiving antenna having a length of 1 metre is of the order of 20 mV.

2. Compare the value obtained above for  $K$  with the value of the solar constant, i.e., the energy that would arrive from the Sun each second on  $1 \text{ cm}^2$  of the Earth's surface if there were no losses in the atmosphere;

$$K_{Sun} = 0.15 \text{ W/cm}^2 = 1,500 \text{ W/m}^2.$$

#### Sec. 119. MOMENTUM AND PRESSURE OF AN ELECTROMAGNETIC FIELD

According to the theory of relativity (see p. 314), matter which possesses energy also possesses mass. The relationship between mass and energy is given by the equation  $E = mc^2$ , where  $c$  is the propagation velocity of light. As we already know, the energy of an electromagnetic field may be considered to have the following density distribution in space:

$$w = \frac{1}{8\pi} (\epsilon E^2 + \mu H^2).$$

Thus, a unit volume of electromagnetic field possesses a mass of  $m = \frac{w}{c^2}$ .

Since moving matter possesses mass, it must also have a momentum equal to the product of the mass and the velocity of motion. We conclude, therefore, that a unit volume of electromagnetic field has a momentum

$$g = mc = \frac{w}{c}.$$

This expression is appropriately called momentum density.

As stated earlier (p. 239), since Poynting's vector has the significance of energy flow, it must be related to the energy density in accordance with the formula  $K = wc$ . Comparing the last two formulas, we see that the relationship between the momentum density and Poynting's vector is given by the expression  $g = K/c^2$ , where  $c$  is the velocity.

Since a flow of electromagnetic radiation possesses mass and momentum, it will exert pressure on a surface placed in its path. The magnitude of this pressure may be expressed in terms of the momentum density and may vary depending on

whether the surface absorbs or reflects the wave energy. Of course, intermediate cases are also possible.

In the time  $\Delta t$ , the electromagnetic field included in a volume  $Sc \Delta t$  strikes the surface  $S$ . If total absorption occurs, a momentum equal to  $gSc \Delta t$  is lost in this time. But momentum divided by time is force, and force divided by area is pressure. Hence, the pressure exerted on the surface absorbing electromagnetic energy is equal to  $p = gc$ , the product of the momentum density and the velocity of light, or, since  $g = \frac{w}{c}$ , the pressure is equal to the energy density  $w$ .

Now, let us consider an ideal elastic encounter between the field and the surface. If all the energy of the electromagnetic field (wave) is reflected, the change in momentum will be twice the incident momentum, for the latter has reversed its direction. Just as in the purely mechanical cases (p. 239), the force of an elastic impact is twice as large as the force of an inelastic impact. Hence, the pressure exerted by the wave on an ideally reflecting plate is

$$p = 2gc \quad \text{or} \quad p = 2w$$

The formula for the general case is now easily obtained. If the plate reflects part of the energy and the coefficient of reflection is equal to  $\rho$ , the pressure of the electromagnetic flux (wave) is given by the expression

$$p = w(1 - \rho) + 2\rho w = (1 + \rho)w.$$

Using light, P. N. Lebedev verified these formulas experimentally in 1900 and thus greatly contributed toward the development of our present conception of the nature of electromagnetic waves. The pressure of light is exceedingly small even for the most intense sources. For example, the pressure of light on a mirror located at a distance of 1 metre from a "lamp" of 1 million candle power is of the order of  $10^{-4}$  dyne/cm<sup>2</sup>. That is why Lebedev's measurement of the pressure of light with an accuracy of 1-2% is viewed as a great experimental achievement.

Basically, Lebedev's apparatus consisted of a pair of vanes attached to a light-weight suspension. One vane was an excellent absorber of light and the other an excellent reflector. The light was directed first at one vane and then at the other, and, then, by comparing the displacement angles, it was possible to determine the magnitude of the force. The chief difficulty was how to take into account the effect on the vanes of residual gas heating in the vessel containing the suspension.

As we have just seen, the theory of variable electromagnetic fields has led to the conception of a field as a physical reality (electromagnetic radiation). The great merit of Lebedev's experiments is that they provided direct proof of the validity of this conception.

An electromagnetic field possesses energy and momentum, is propagated in space with a specific velocity and exerts pressure on an obstacle. We shall see below that an electromagnetic field may be transformed into matter. All these facts taken together irrefutably prove that an electromagnetic field is a physical reality.

# Electromagnetic Radiation

## Sec. 120. ELEMENTARY DIPOLE

Electromagnetic radiation occurs whenever a variable electromagnetic field is created in space. An electromagnetic field, in turn, varies in time whenever the distribution of electric charge in a system changes or the density of an electric current varies.

Thus, every variable current and pulsating electric charge is a source of electromagnetic radiation.

Magnetic and electric dipoles having variable moments—particularly the latter—are the simplest systems producing electromagnetic fields. A system consisting of a stationary positive charge about which a negative charge oscillates constitutes such an electric dipole. If the oscillation is sinusoidal, the dipole moment will also be sinusoidal, i.e., it is represented by the formula  $p = p_0 \cos \omega t$ . This simple radiator model has very great significance since many real systems can be represented to a high degree of accuracy by ideal dipoles.

It will be recalled (see Sec. 93) that the electric properties of a system whose “centres of gravity” of positive and negative charge do not coincide may be described in terms of the dipole moment of the system. But most radiators of electromagnetic energy are electrically neutral systems whose positive and negative charges are capable of being displaced relative to each other. This is primarily because atomic and molecular systems fall under this heading. An electron rotating about the nucleus of an atom is a system having a variable dipole moment, and a neutral molecule whose atoms are in a state of oscillation is also, frequently, a system having a variable dipole moment. However, our interest in the electric dipole extends further. In the following article, we shall see that a linear radio antenna may be likened to a dipole. (Incidentally, the analogous terms “oscillator” and “vibrator” are somewhat broader in meaning than the exact term “dipole”.)

Magnetic dipoles occur when the electric charge distribution and hence the dipole moment of the system remain unchanged while the current density and hence the magnetic moment of the system change with time. A typical example is a loop in which an alternating electric current flows. If the current flows in a closed circuit, the electric charge is neither accumulated nor dissipated anywhere. The electric dipole moment of such a loop equals zero and does not change. However, the loop's magnetic field, which is related to the value of its magnetic moment, varies and, therefore, electromagnetic energy is radiated. It follows from the theory that if a system possesses simultaneously an electric and a magnetic moment the radiation from the magnetic dipole at large distances from the source is usually much less than the radiation from the electric dipole.

If a dipole radiates by giving up internal energy or, as in the case of an antenna, by transforming the energy of an external source into radiation energy, the dipole is called a *primary radiator*. However, a *secondary radiator* is also of considerable interest. In this case, a dipole is made to oscillate by the action of an electromagnetic wave and becomes a radiator only as a consequence of this action. Secondary oscillations are particularly intense when the primary wave is of the same frequency as the natural frequency of the dipole (resonance).

Setting a dipole into an oscillatory state may be viewed as a mechanical process—the jostling of the charges by an external force equal to the product of the charge

and the field intensity. At the same time, the process of creating secondary oscillations in a receiving antenna may be viewed as an induction process in which an alternating electric current is produced by an alternating magnetic field. To the extent that the antenna may be replaced by a dipole, both views are equivalent.

#### Sec. 121. ANTENNAS AS ELECTRIC DIPOLES

An important difference exists between the state of oscillation of an oscillatory circuit (p. 236) and the oscillation of the current in an antenna. In discussing the electric oscillation of a circuit, we referred to a definite instantaneous current strength and a definite instantaneous charge on the condenser plates. It was assumed that the current strength in all parts of the circuit was the same, that the electric charge was concentrated on the condenser plates and, hence, at a given instant could only have a single value.

The electric oscillation in the case of an antenna cannot be viewed in the same manner as the oscillation of a pendulum. However, the oscillation of an electric

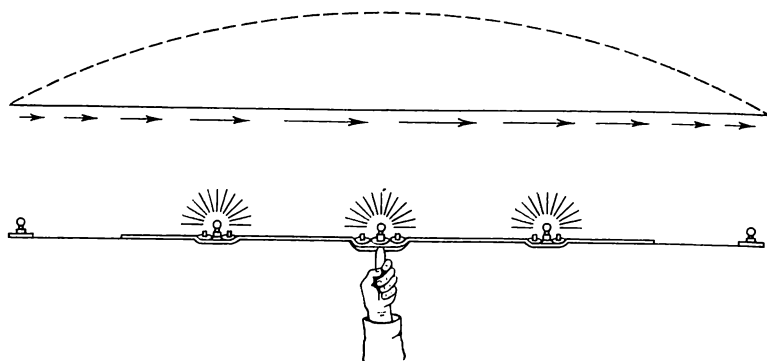


Fig. 132

current in an antenna does have a mechanical analogue. This oscillation is very similar to the vibration of a rod or string, i.e., it can be represented by a standing wave.

This may be strikingly demonstrated by showing that an excited antenna has current nodes and antinodes. A small bulb may serve as the current indicator (Fig. 132). It turns out that a conduction current antinode exists at the centre of a free section of wire in which electromagnetic waves are excited and that conduction current nodes exist at the ends. In such a wire, the current at all points is unidirectional at each instant. At some instant the current at each point decreases to zero and then begins to flow in the opposite direction. The electric charge, which is distributed continuously along the wire, varies accordingly. Clearly, as long as current flows in one direction, positive charge is accumulated on one half of the wire and negative charge is formed on the other. When the current decreases to zero, the charges at the ends are a maximum and of opposite sign. The current then begins to flow in the opposite direction and the charges decrease, becoming zero when the current strength in all parts of the line is a maximum. At this instant, the recharging process commences, charges of opposite sign are accumulated on the two halves of the wire, etc.

The reader will note that at each instant charges of opposite sign are located on the two halves of the wire. Two charges that are equal in magnitude but opposite in sign, and separated by a certain distance, constitute an electric dipole. It can be stated, therefore, that the electric oscillations of an antenna are very similar to the oscillations of an electric dipole in which the dipole moment decreases from a maximum positive value to zero, then increases in the opposite direction, then again decreases, etc.

The field of an antenna differs from that of a dipole only in the region close to the antenna. At distances hundreds of times greater than the dimensions of the antenna, the field created by the antenna does not differ from the field created by an ideal electric dipole.

Let us again return to the analogy between an antenna and a rod. The natural frequencies of electric oscillations that can exist in free, ungrounded antennas are not restricted to the frequencies determined from the simplest case, i.e., when a half wavelength is impressed on the length of the antenna, although such half-wavelength dipoles are mainly used in UHF engineering. The following relationship exists between the length of the antenna and the wavelength:  $L = n \frac{\lambda}{2}$  (see p. 98). Thus, an antenna of length  $L$  can receive and radiate waves of wavelength  $\lambda$  satisfying the above relationship.

In the field of radio, a number of methods exist for varying the natural frequencies of an antenna. Basically, they consist in the connection of a self-inductance coil or a condenser to the antenna. By varying the inductance or capacitance, the natural frequencies of the antenna may be varied within broad limits.

#### Sec. 122. RADIATION PATTERN OF A DIPOLE

The radiation pattern of a dipole may be determined experimentally. It turns out that the results are in complete accord with the theory first advanced by Hertz. We shall be concerned only with the results of experiments and theoretical calculations, restricting ourselves to the field far removed from the dipole, i.e., to the so-called wave zone. This is the region in which the distances to the dipole are considerably greater than the dipole dimensions.

Irrespective of the complexity of dipole oscillations, the oscillations may always be resolved by means of Fourier's theorem into their spectra, i.e., they may be represented as the sum of harmonic oscillations of frequencies  $\omega$ ,  $2\omega$ ,  $3\omega$ , etc. Therefore, it is quite sufficient to consider the electromagnetic field of a dipole whose moment varies in accordance with the harmonic relation  $p = p_0 \cos \omega t$ .

Calculations and experiments show that the field of such a system may be represented by a spherical wave propagating with a velocity  $v = \frac{c}{\sqrt{\epsilon\mu}}$ . The electric and magnetic vectors of the wave are at right angles to each other and also at right angles to the direction of propagation. The latter circumstance, incidentally, follows from Poynting's theorem.

In the wave zone, the electric and magnetic vectors vary in phase, performing harmonic oscillations at every point in space. A simple relationship exists between the numerical values of the field intensity vectors, namely:

$$\sqrt{\epsilon E} = \sqrt{\mu H}.$$

Hence, Poynting's vector may be written in the following form:

$$K = \frac{c}{4\pi} E^2 \sqrt{\frac{\epsilon}{\mu}} = \frac{v}{4\pi} \epsilon E^2.$$

Thus, the wave intensity, i.e., the energy passing through unit area per unit time, is proportional to the amplitude squared of the electric field intensity.

The radiation of a dipole is not the same in all directions. The amplitude, as well as the intensity, depends on the angle of inclination of the propagation direction to the axis of the dipole. In the direction perpendicular to the dipole axis the

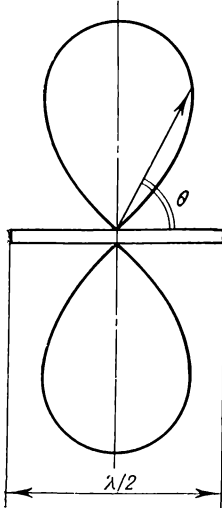


Fig. 133

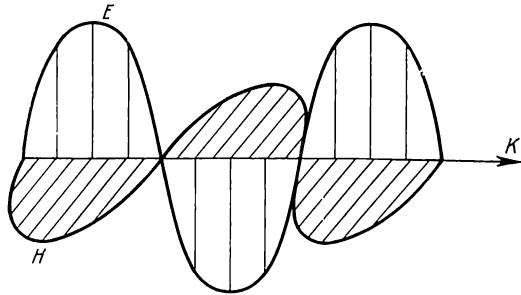


Fig. 134

radiation is a maximum, while in the direction of the dipole moment it is equal to zero. Theory gives us the following expression for the electric field intensity:

$$E = \frac{p_0 \omega^2}{c^2 R} \sin \theta \cos \omega \left( t - \frac{R}{v} \right),$$

where the factor before the cosine is the wave amplitude of vector  $E$ . The angle  $\theta$  is the angle between the propagation direction and the dipole axis. The expression for the magnetic field deviates from the above only with respect to a slight difference in the amplitude factor.

Fig. 133 shows a diagram sometimes used to represent the dependence of radiation intensity on direction. Here, a radius vector is shown intersecting the radiation pattern. If the scale is known, the radiation intensity is given by the length of the vector measured to the point of intersection.

The fact that the amplitude is proportional to the radiation frequency squared is very important. Clearly, the radiation intensity of the dipole depends very greatly on the frequency, i.e., it is proportional to the frequency to the fourth power:

$$K \sim \frac{\omega^4}{R^2} \sin^2 \theta.$$

Thus, when the frequency is halved, the intensity decreases to  $\frac{1}{16}$  of its original value.

Theory has led to an important conclusion regarding the orthogonality of an electromagnetic wave. This is illustrated in Fig. 134, where it is seen that the electric and magnetic vectors are perpendicular to the direction of propagation. As a result, the properties of an electromagnetic wave change when the wave is turned



about the direction of propagation. This phenomenon is known as *polarisation*. Mapping the flux lines of a radiating dipole is of no particular interest. Fig. 135 shows vectors of electric field intensity for several points in space. The field is rotational and the flux lines are closed. When radiation occurs, the close lines expand in the direction away from the radiator. The magnetic lines of force consist of circles about the dipole axis. At great distances from the dipole, the spherical wave practically does not differ from a plane wave. Of course, the orientation of the  $E$ ,  $H$  and  $K$  vectors in a plane wave and the numerical relations given above remain the same.

### Sec. 123. THE ELECTROMAGNETIC SPECTRUM

According to theory, electromagnetic radiation occurs when electric charges are accelerated non-uniformly. A uniform or free flow of electric charge does not produce radiation. Charges moving under the action of a constant force, e.g., charges describing a circle in a magnetic field, also do not radiate.

In oscillatory motion, the acceleration is continuously changing. Hence, electric charge oscillations produce electromagnetic radiation. Electromagnetic radiation also occurs when charges are abruptly decelerated. Thus, when a beam of electrons impinges on a target, X-rays are produced. Electromagnetic radiation also occurs during random thermal motion of particles (thermal radiation). The pulsations of a nuclear charge produce an electromagnetic radiation known as  $\gamma$ -rays. Ultraviolet rays and visible light are produced by the motion of atomic electrons. Electric charge oscillation on a cosmic scale is exemplified by the radiation of radio waves by heavenly bodies.

In addition to natural processes in which various kinds of electromagnetic radiation are produced, a number of experimental means exist for creating electromagnetic radiation.

The main characteristic of electromagnetic radiation is its frequency (in the case of a harmonic oscillation) or its frequency band. Of course, using the relation  $c = \nu\lambda$ , the length of the electromagnetic wave in vacuum may be determined if the radiation frequency is known.

The radiation intensity is proportional to the frequency to the fourth power. Hence, very low-frequency radiation having wavelengths of the order of hundreds of kilometres cannot be traced. Practically, the radio band begins with wavelengths of the order of 1 or 2 km, which corresponds to frequencies of the order of 150 kHz. Wavelengths of the order of 200 metres are in the medium-frequency band, while those of the order of tens of metres are in the short-wave band. Ultrahigh frequencies (UHF) are beyond the usual radio frequencies; wavelengths of the order of several metres and fractions of a metre, down to a centimetre (i.e., frequencies of the order of  $10^{10}$ - $10^{11}$  MHz), are used in the television field and in radar.

In 1924, Glagolyeva-Arkadyeva obtained even shorter electromagnetic waves. Electric sparks produced between iron filings suspended in oil served as her source. In this manner, wavelengths down to 0.1 mm were obtained. Thus, an overlap was achieved with thermal radiation wavelengths.

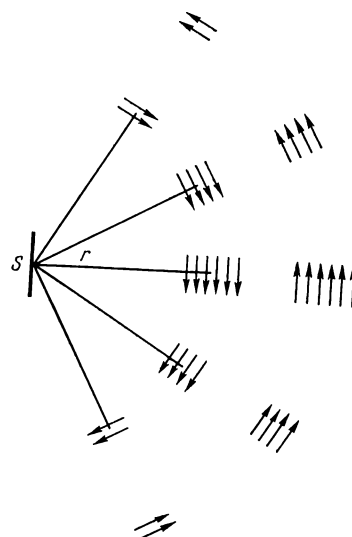


Fig. 135

Visible light occupies a very small band of wavelengths—from  $7.6 \times 10^{-5}$  cm to  $4 \times 10^{-5}$  cm. This band is followed by ultraviolet rays, which are invisible but very easily detected by means of laboratory equipment. These wavelengths extend from  $4 \times 10^{-5}$  cm to  $10^{-5}$  cm.

Following the ultraviolet band is the X-ray band. The wavelengths in this band extend from  $10^{-6}$  cm to  $10^{-10}$  cm. The shorter the X-ray wavelength, the less the absorption by matter. Electromagnetic radiations of shortest wavelength, which are the most penetrating, are called  $\gamma$ -rays (wavelengths of  $10^{-9}$  cm and less).

The nature of any of the enumerated electromagnetic radiations may be completely determined as follows: first, the electromagnetic radiation is resolved into a spectrum by some method or other. In the case of light, ultraviolet rays and infrared radiation, this may be accomplished by refraction through a prism or by passing the radiation through a diffraction grating (see below). In the case of X-rays and  $\gamma$ -rays, resolution into a spectrum is achieved by reflection from a crystal (see p. 293). The spectrum of radio waves is determined by making use of the phenomenon of resonance.

The radiation spectrum obtained may be continuous or discrete, i.e., all frequencies may be present in a broad band of the radiation spectrum or the spectrum may consist of individual sharp lines corresponding to very narrow bands of frequency. In the first case the spectrum is represented by a curve of intensity vs. frequency (or wavelength), while in the second case the spectrum is described by giving the frequency and intensity of the lines.

Experiments show that electromagnetic radiation of given frequency and intensity may not always have the same polarisation state. In addition to radiation in which the electric vector of the waves oscillates along a specific line (linearly polarised waves), radiation also exists in which linearly polarised waves turned relative to each other about the beam axes are superimposed. Hence, to completely describe radiation, it is also necessary to indicate its polarisation.

It should be noted that even for the slowest electromagnetic oscillations, the electric and magnetic vectors of a wave cannot be measured. The above descriptions of a field are based on theory. Nevertheless, in view of the continuity and unity of all electromagnetic theory, there is no reason to doubt their veracity.

The assertion that one or another kind of radiation consists of electromagnetic waves is always based on indirect evidence. However, since these hypotheses have so many consequences that are in complete mutual agreement, the electromagnetic spectrum hypothesis long ago became accepted fact.

#### Sec. 124. QUANTUM NATURE OF RADIATION

We have already noted (p. 115) that investigation of atomic phenomena has led to a law which states that the internal energy of a system cannot assume any arbitrary value, but is characterised instead by a system of energy levels. Energy radiation is related to the transition of a system from a higher level to a lower level. Energy absorption is related to the transition to a higher level.

This applies, in the first place, to electromagnetic radiation. The quantum nature of submicroscopic phenomena was discovered at the beginning of this century as a result of investigation of a number of conflicting facts regarding electromagnetic radiation.

Thus, the emission of electromagnetic radiation of frequency  $\nu$  by a dipole occurs in quanta (packets) rather than continuously. A quantum of energy is equal to  $h\nu$ , where  $h$  is Planck's constant and is equal to  $6.62 \times 10^{-27}$  erg sec =  $6.62 \times 10^{-34}$  J sec.

The quantum nature of electromagnetic waves is manifested in absorption as well as radiation, for absorption too can only occur by means of energy quanta. If the value of an energy quantum is equal to the difference between certain energy levels of a system on which the wave impinges, the absorption process is quite pronounced. Such a process may be called resonance absorption. From the standpoint of classical physics, such absorption occurs when the frequency of the external field is equal to the oscillation frequency of the particles constituting the system. If the value of the electromagnetic wave quantum is less than the difference between energy levels, absorption cannot occur and the wave passes freely through the system.

In quantum terms, the secondary radiation of a system is described as follows: a system absorbs a quantum of electromagnetic energy and is raised to a higher energy level. The system maintains this level for a certain period of time and then returns to its former energy level by giving up energy—again in the form of a quantum.

Since a quantum of energy is equal to  $h\nu$ , it is immediately evident that the higher the radiation frequency, the more pronounced the quantum phenomena. Nevertheless, the quantum nature of radiation has already been observed in practically all regions of the electromagnetic spectrum. It has even been possible to observe the quantum absorption of radio waves having wavelengths of several hundred metres.

The presence of one or another spectrum of electromagnetic radiation depends, in the first place, on the arrangement of energy levels in the system under consideration and on the transition probabilities of the system from an  $n$ -th level to an  $m$ -th level. If these probabilities were known beforehand and the energy level diagram were available, it would be an easy task to determine the radiation spectrum of the system.

We shall repeatedly be dealing with problems of radiation and absorption of electromagnetic energy, but now let us consider some problems of electromagnetic wave propagation in which the quantum nature of radiation is not manifested when the phenomenon is not accompanied by the absorption and radiation of energy.

# Propagation of Electromagnetic Waves

## Sec. 125. DISPERSION AND ABSORPTION

In a homogeneous medium, the velocity of propagation and the direction of an electromagnetic wave do not change. The velocity of the wave is a maximum in vacuum. In a medium, the wave velocity is

$$v = \frac{c}{\sqrt{\epsilon\mu}},$$

and since in most practical cases  $\mu = 1$ , we obtain

$$v = \frac{c}{\sqrt{\epsilon}}.$$

The ratio of the velocity of wave propagation in vacuum to the velocity of propagation in a medium is called *the index of refraction*. Thus, from electromagnetic theory we obtain the equation  $n = \sqrt{\epsilon}$ , which is quite valid for very long wavelengths. As the wavelength changes, the index of refraction changes. This *dispersion* is alien to Maxwell's electromagnetic theory, which regards a medium as a continuum and does not take into account the interaction of radiation and matter. Be that as it may, the equation  $n = \sqrt{\epsilon}$  is not valid for rapid electromagnetic oscillations.

When an electromagnetic wave is propagated through matter, the electric charges of the molecules are set into a vibratory state. Since an electron cloud moves freely as compared with heavy nuclei, electric oscillation consists in the displacement of the centre of gravity of the electrons relative to the stationary centre of gravity of the atomic nuclei's positive charges. Designating the charge and mass of the oscillating electrons by  $e$  and  $m$ , respectively, the oscillation equation may be written in the form

$$m\ddot{x} = -kx - eE_0 \cos \omega t$$

or, dividing by  $m$  and using the formula for the natural frequency of oscillation, i.e.,  $\omega_0^2 = \frac{k}{m}$ , we obtain

$$\ddot{x} = -\omega_0^2 x - \frac{e}{m} E_0 \cos \omega t.$$

We have equated the product of mass and acceleration to two forces—the restoring force— $kx$  and the external, periodically varying force  $eE_0 \cos \omega t$ . This is the equation of forced harmonic oscillations. It is satisfied if

$$x = x_0 \cos \omega t.$$

After substituting in the equation, we obtain

$$x_0 = \frac{-\frac{e}{m} E_0}{\omega_0^2 - \omega^2}.$$

The dipole moment of a molecule is

$$ex_0 = \frac{\frac{e^2}{m} E_0}{\omega_0^2 - \omega^2}.$$

The polarisation vector, i.e., the dipole moment per unit volume, is  $N$  times larger, where  $N$  is the number of molecules per unit volume:

$$P = \frac{\frac{Ne^2}{m}}{\omega_0^2 - \omega^2} E.$$

Recalling the formula relating the polarisation to the field intensity, i.e.,

$$P = \frac{\varepsilon - 1}{4\pi} E,$$

it is seen that the permittivity of the medium has been expressed in terms of the parameters of the molecular dipole:

$$\varepsilon = 1 + \frac{\frac{4\pi Ne^2}{m}}{\omega_0^2 - \omega^2}.$$

The index of refraction of the medium should be equal to the square root of this expression.

As shown in Fig. 136, the nature of the dependence is confirmed by experiment. Here, the index of refraction vs. frequency curve, based on the above formula, is compared with the curve based on measurements with specific substances.\* What is the basic conclusion to be drawn from the experimental and theoretical results? In general, the index of refraction increases with increasing frequency in the entire frequency range, except for the region in the immediate vicinity of resonance absorption. This region is called the *anomalous dispersion* region. A substance may have several resonance frequencies, which correspond to the differences between its energy levels. Hence, there will be a corresponding number of anomalous dispersion regions.

Thus, the index of refraction of a wave, and hence the velocity of propagation, greatly depends on the value of the wave frequency relative to the natural frequencies of the molecular dipoles.

Naturally, the capacity of a substance to absorb an electromagnetic wave depends on the same factors. Using the same reasoning as in the case of elastic waves (see p. 89), we arrive at a completely analogous formula:

$$I = I_0 e^{-\mu d},$$

which enables us to determine the value of the radiation intensity  $I$  relative to the incident intensity  $I_0$  if the absorption coefficient  $\mu$  and the layer thickness  $d$

\* A more exact theory results in agreement with the experimental values in the region close to  $\omega_0$  as well.

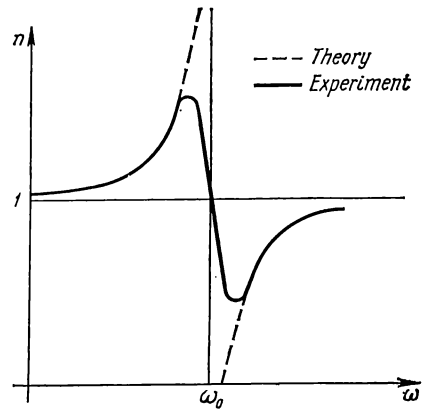


Fig. 136

through which the wave has passed are known. It should be recalled that the absorption coefficient is equal to the reciprocal of the layer thickness which decreases the radiation intensity to  $\frac{1}{e}$  of its original value. Due to the complex system of energy levels peculiar to matter, the curve of the absorption coefficient plotted against the frequency of the incident wave may appear odd and "erratic".

Until now we have been considering dielectric media containing only bound electric charges. Other relations exist when an electromagnetic wave is propagated in a medium in which a considerable number of free electrons are present. Such media include metals and the ionosphere—a region of free charges akin to a gas. Using the theory presented above, it must be assumed that in the formula for  $\epsilon$  the natural frequency  $\omega_0$  of a free charge is equal to zero (the frequency is proportional to the rigidity of the bond). The dielectric constant is then given by the formula

$$\epsilon = 1 - \frac{4\pi N e^2}{\omega^2} \cdot m$$

When  $\omega$  becomes sufficiently large, the index of refraction  $n = \sqrt{\epsilon}$  approaches unity. But when  $\omega^2 < 4\pi N e^2/m$ , the index of refraction is imaginary. This means that for the given values of frequency the waves cannot penetrate the metal or the ionosphere. On the other hand, for high frequencies the waves are "indifferent" to the presence of a medium containing electrons. These predictions are borne out in the case of radio waves. Thus, long and medium waves are reflected from the ionosphere and do not penetrate it, short waves penetrate the ionosphere and UHF waves pass through the ionosphere unimpeded.

The above presentation is greatly oversimplified. Hence, it should not be surprising that the conclusions are not valid for the optical region where the values of the index of refraction may be close to zero and also much greater than unity.

#### Sec. 126. BEHAVIOUR OF AN ELECTROMAGNETIC WAVE AT THE BOUNDARY BETWEEN TWO MEDIA

Just as in the case of an elastic wave, an electromagnetic wave is reflected and refracted at the boundary between two media. The basic laws of these phenomena

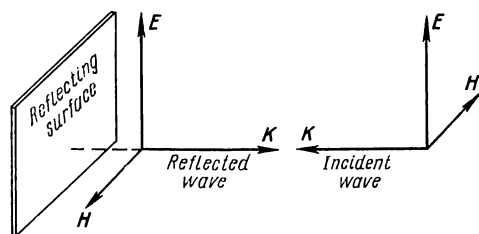


Fig. 137

may be subjected to theoretical analysis by utilising the boundary conditions on the electromagnetic field vectors. These conditions, discussed on pp. 195 and 219, follow in turn from Maxwell's equations. Since the relationships between the fields on either side of the boundary are not arbitrary, the division of the wave into reflected and transmitted components is also not arbitrary.

The two fundamental relations may be expressed as follows: the tangential components of the electric and magnetic vectors on either side of a boundary must be equal.

What restrictions are imposed by these relations in the simple case of normal incidence? This case is illustrated in Fig. 137. Assuming that the electric vectors are in the plane of the page, then the magnetic vectors are perpendicular to this plane. We know that the electric and magnetic vectors and the direction of propa-

gation may be viewed as a right-handed screw system, i.e., the rotation of vector  $E$  by the shortest path toward vector  $H$  appears counterclockwise to one facing the oncoming wave. To satisfy this requirement of electromagnetic theory, the direction of either vector  $H$  or vector  $E$  must be reversed for the reflected wave. Thus, either the magnetic or the electric vector undergoes a  $180^\circ$  phase shift when the wave is reflected.

When considering oblique incidence, it is necessary to determine which of the two actually occurs. It turns out that both are possible—one when the wave passes into a medium of larger  $\epsilon$  and the other when it passes into a medium of smaller  $\epsilon$ .

For normal incidence, the following calculations do not depend on the scheme chosen. Let us write the boundary conditions in the form

$$E_{\text{incid}} = E_{\text{reflect}} + E_{\text{refract}}$$

and

$$H_{\text{incid}} = -H_{\text{reflect}} + H_{\text{refract}}.$$

But the following relationship exists between the numerical values of the  $H$  and  $E$  vectors:

$$H = \sqrt{\epsilon} E = nE.$$

Hence, we obtain two equations,

$$E_{\text{incid}} = E_{\text{reflect}} + E_{\text{refract}}$$

and

$$n_1 E_{\text{incid}} = -n_1 E_{\text{reflect}} + n_2 E_{\text{refract}}$$

whence the ratios  $E_{\text{reflect}}/E_{\text{incid}}$  and  $E_{\text{refract}}/E_{\text{incid}}$  may be determined. Since the wave intensity is proportional to the amplitude squared and the index of refraction (p. 246), we obtain for the coefficients of reflection and transmission the following simple formulas, where  $n = \frac{n_2}{n_1}$  is the relative index:

$$\text{coefficient of reflection} = \left( \frac{E_{\text{reflect}}}{E_{\text{incid}}} \right)^2 = \frac{(n-1)^2}{(n+1)^2},$$

$$\text{coefficient of transmission} = \left( \frac{E_{\text{refract}}}{E_{\text{incid}}} \right)^2 = \frac{4n}{(n+1)^2}.$$

The similarity to the case of elastic waves is very great.

By means of such calculations, the general results presented below were obtained for the case of arbitrary beam inclination and polarisation state of the wave. The agreement with experimental results is quite satisfactory.

Since the sum of the reflection and transmission coefficients is equal to unity, the theoretical results are completely described by Fig. 138, where the intensity of the reflected wave is plotted as a function of the angle of incidence.

Calculations and experiments show that the nature of the reflection curve depends to a significant extent on the polarisation state of the incident wave relative to the plane of incidence. The electric field intensity vector  $E$  is "more important" than the  $H$  vector, if only in the sense that the photochemical action is due to  $E$ . Therefore, in describing the polarisation state of a wave, it is customary to specify it with respect to the electric vector. The orientation of the  $H$  vector is always easily found if the direction of propagation is known. Thus, it turns out that the reflection coefficient is different for two waves that are incident at one and the same

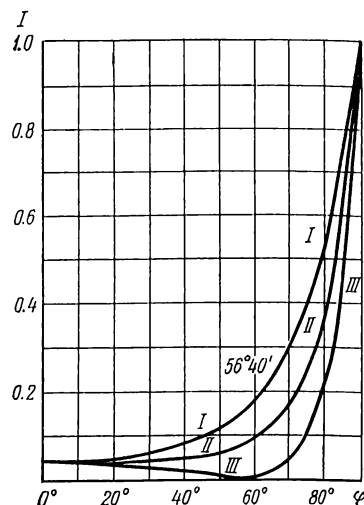


Fig. 138

angle  $\varphi$  on the same boundary if in one case the electric vector is in the plane of incidence and in the other it is perpendicular to this plane. In the figure, curve *I* corresponds to the case when the  $E$  vector is perpendicular to the plane of incidence, curve *III* corresponds to the case when the  $E$  vector is in the plane of incidence and curve *II* corresponds to the case when the wave is not polarised.

In the first case, the change in the reflection coefficient is monotonic—for normal incidence there is little reflection, the coefficient being of the order of 5%; then, with increasing angle the reflection coefficient increases, ever more rapidly, until the glancing angle is reached. A beam whose electric vector is in the plane of incidence behaves in an entirely different manner. Its reflection intensity decreases with increasing angle until it reaches zero at the angle  $\varphi_B$ . This angle is determined by the following interesting equation:  $n = \tan \varphi_B$ . The figure is plotted for the value  $n = 1.52$  (transition from air to glass). Hence, the angle at which the reflection coefficient decreases to zero is equal to  $56^\circ 40'$ . For a further increase in angle, the coefficient of reflection begins to increase and finally reaches unity.

What is the reason for the absence of reflection in this particular case? How does this case differ from others? Evidently, the answer must be sought in the boundary conditions, from which the entire theory of the phenomenon proceeds. We leave it to the reader to construct the field vectors for this angle and illustrate the requirement.

The following question may arise in the reader's mind: if the boundary conditions enable us to understand all phenomena at the boundary between two media, then what about total internal reflection where there is a field in one medium ( $E_{1t} \neq 0$ ) but none in the other? The question is perfectly valid and the theory provides an answer. It turns out that under conditions of total internal reflection the field penetrates the second medium but is not propagated deeply into the medium. The condition  $E_{1t} = E_{2t}$  is not violated.

A number of experiments have been devised for the demonstration of light wave penetration into a second medium under conditions of total internal reflection. Suffice it to recall the basically simple experiment proposed by Mandelshtam. A glass prism is partially immersed in a solution of fluorescein—a substance exhibiting a characteristic fluorescence under the action of light. Then, a beam of light is directed onto the prism in such a manner that total reflection occurs on the inner side of the prism surface that is immersed in the solution. The fluorescein thereupon glows intensely in an extremely thin layer next to the glass, proving that the electromagnetic wave has penetrated the solution.

#### Sec. 127. NATURAL AND POLARISED LIGHT. POLARISATION UPON REFLECTION

Place a glass plate at an angle  $\varphi_B$  to a light beam. The beam is reflected. Then, if the beam is turned about its axis (actually, the source of light is turned about the beam axis) it might be expected that at some position the beam will not be reflected. But if natural light is used for the experiment, this does not occur, i.e., for every azimuthal position of the incident beam, the reflected beam has the same intensity. It would be wrong to consider this a refutation of the theory presented in the preceding article. This experiment merely shows that the polarised state of a beam of natural light is more complex than given by the scheme of two vectors,  $E$  and  $H$ , having fixed directions of oscillation.

Now, let the above beam, reflected at an angle  $\varphi_B$ , impinge on a second plate placed at a similar angle  $\varphi_B$  to the beam reflected from the first plate. Then, turn the beam about its axis. Since of course, only the relative position of the beam and the reflector is of importance, it is easier to turn the second glass plate. Investiga-



tion of this double reflection shows that the reflection varies with the position, and the position for which no reflection occurs is easily found. It is evident that this position corresponds to a mutual orientation of beam and reflector for which the electric vector of the beam is in the plane of incidence. The following conclusion may be drawn: reflection from the first reflector results in the natural beam acquiring a polarised state in which a single oscillation direction of the electric vector is separated out.

In contradistinction to natural beams, beams in which the vectors have a specified oscillation direction are referred to as polarised beams. How should the polarised state of a natural beam be envisaged? It is necessary to assume that in a natural electromagnetic wave all possible oscillation directions of the electric vector are uniformly present. The word "possible" should be underlined since electromagnetic theory shows that the electric vector is of a transverse nature. In essence, therefore, a natural unpolarised wave is a superposition of numerous linearly polarised waves having a uniform distribution as regards the vector oscillation directions. All transverse directions are electric vector oscillation directions of a beam of natural light.

Reflection from two successive reflectors, fixed at an angle  $\varphi_B$  to the beams, is one method of polarising beams of light.

An electric vector of natural light may always be resolved into two mutually perpendicular components. When reflection is being investigated, it is most convenient to resolve each vector into two components—one in the plane of incidence and the other perpendicular to this plane. Thus, the behaviour of a natural beam may be equated to the behaviour of two such component waves, if we take into account that the phase difference between them varies randomly. Therefore, in describing the polarisation of light, we say that one of the components has not been transmitted, or has been transmitted to such and such an extent. If upon reflection or refraction one of the components of light is transmitted to a larger extent than the other—which the reflection curves show to be the case—this signifies that the light has been partially polarised.

We can utilise this phenomenon to obtain total polarisation of a beam. Instead of using two reflectors fixed at an angle  $\varphi_B$  to the beams, it is much easier to transmit a beam through a pack of glass plates. Each refraction will increase the share of one of the components in the beam by a certain percentage. In this manner almost total polarisation may be achieved.

The natural state of a light beam is unpolarised. However, this does not mean that every beam that has not been subjected to reflection or refraction is unpolarised. This applies particularly to radio waves. The short electromagnetic waves used in the transmission of television are highly polarised. It is precisely this circumstance that enables us to determine the direction of the transmitter by the orientation of the receiving antenna. The electromagnetic waves which act as the carrier of a television programme are highly polarised. Hence, the antenna must be oriented in such a manner that the oscillation direction of the electric vector coincides with the antenna direction.

#### Sec. 128. PROPAGATION OF LIGHT WAVES IN A MEDIUM HAVING A REFRACTIVE INDEX GRADIENT

As a rule, a difference in density is associated with a difference in refractive index. The natural question arises: what is the nature of the wave propagation in a medium in which the value of the refractive index varies from point to point, i.e., the refractive index gradient differs from zero?

A difference in refractive index signifies a difference in the velocity with which the wave front advances. It follows, therefore, that the wave front will be continuously deformed as it advances in such a medium. If we construct the normals to the wave front, we obtain a curved line. Thus, it can be stated that in a nonhomogeneous medium light is propagated curvilinearly rather than rectilinearly.

The analogous problem for sound waves was discussed earlier (p. 110). The same laws are applicable here and the beam path is also determined by Fermat's principle. A beam of light propagated in a finite medium having a refractive index gradient follows a path between two points that requires a minimum amount of time to traverse. Therefore, the beam of light bends so as to shorten its path in regions where the refractive index is large and lengthen its path in regions where the refractive index is small.

The best example of the propagation of light in a medium of gradient  $n$  is the passage of a beam of light through the Earth's atmosphere. Since the density and the index of refraction of air decrease with increasing elevation, it follows that refraction occurs in the atmosphere. A beam travelling from a star to the Earth and entering the atmosphere at an angle rather than along a radius will bend; hence, the apparent position of the star is displaced relative to its true position. For a star at the zenith, the displacement angle is as much as  $1/2$  of a degree.

Mirages are caused by the presence of a refractive index gradient in the atmosphere. They occur in the African deserts due to the fact that heat currents are easily formed above the hot sand, resulting in temperature gradients and, hence, density and refractive index gradients. As a result, the light beams travel along curved lines and a landscape seems to appear where the observer, accustomed to the rectilinear propagation of light, conceives it.

Of course, in the case of light propagation in a nonhomogeneous medium, the waves are neither spherical nor planar. It should be recalled that a variable propagation velocity signifies that the wavelength also varies from point to point. What is the equation of wave motion in a medium where the refractive index changes from point to point? Since the parameters of the wave change from point to point, the equation being sought must be a differential equation, for only a differential equation can express the dependence between the physical quantities for a given point in space.

This equation may be found by means of Maxwell's equations. Since the derivation is rather complex, it will not be presented here. The calculations yield the following relation, which is valid for the  $E$  vector (or its projection) as well as the  $H$  vector (or its projection):

$$\frac{\partial^2 \Psi}{\partial s^2} = \frac{1}{v^2} \frac{\partial^2 \Psi}{\partial t^2}.$$

The function  $\Psi$  is called *the wave function*. It represents the  $E$  vector, the  $H$  vector, or their components since the equations are the same for each case. Here,  $s$  is the coordinate in the direction of wave propagation,  $t$  is the time and  $v$  is the propagation velocity.

This equation is called *the wave equation* and it is valid for all points in space located outside the sources of the field, i.e., outside charged regions and regions in which electric currents flow.

First it will be shown that the above differential equation is satisfied by the simplest wave process, i.e., a plane wave. As we know (p. 85), the expression for plane wave of frequency  $\omega$ , propagating in the direction  $s$ , has the form

$$\Psi = A \cos \omega \left( t - \frac{s}{v} \right).$$

Let us determine the second derivative of the wave function  $\Psi$  with respect to time and with respect to the coordinate. We obtain

$$\frac{\partial^2 \Psi}{\partial t^2} = -\omega^2 \Psi; \quad \frac{\partial^2 \Psi}{\partial s^2} = -\frac{\omega^2}{v^2} \Psi.$$

It is seen that the following relationship must exist between the second derivatives

$$\frac{\partial^2 \Psi}{\partial s^2} = \frac{1}{v^2} \frac{\partial^2 \Psi}{\partial t^2};$$

hence, the equation of a plane wave is embodied in the proposed differential equation. However, the above differential equation embraces much more. Any function of argument  $\left(t - \frac{s}{v}\right)$  is a solution of the equation since for any function  $\Psi\left(t - \frac{s}{v}\right)$  the derivative expressions in  $\Psi$  are the same.

The dependence of a function on the argument  $\left(t - \frac{s}{v}\right)$  is regarded as the sole indication of a wave process. The significance of this argument consists in the following: if the state at a point  $s = 0$  is characterised at the instant of time  $t = 0$  by a certain value of the wave function, then the same state occurs at point  $s_1$  at the instant of time  $t_1 = \frac{s_1}{v}$ , at point  $s_2$  at the instant of time  $t_2 = \frac{s_2}{v}$ , etc. Here  $s$  is the coordinate reading along any rectilinear or curvilinear path.

The differential equation

$$\frac{\partial^2 \Psi}{\partial s^2} = \frac{1}{v^2} \frac{\partial^2 \Psi}{\partial t^2}$$

is the general equation of a wave process and is valid for any medium, including nonhomogeneous medium where  $v$  varies from point to point.

If it is necessary to express the wave function in terms of the three space coordinates  $x, y, z$ , the generalised wave equation has the following form:

$$\frac{\partial^2 \Psi}{\partial x^2} + \frac{\partial^2 \Psi}{\partial y^2} + \frac{\partial^2 \Psi}{\partial z^2} = \frac{1}{v^2} \frac{\partial^2 \Psi}{\partial t^2}.$$

The sum of the second partial derivatives of a function is concisely designated by the symbol  $\Delta \Psi$  (read: Laplacian of  $\Psi$ ). Thus,

$$\Delta \Psi = \frac{1}{v^2} \frac{\partial^2 \Psi}{\partial t^2}.$$

The differential equation of a wave is valid for any process in which the value of the wavelength and the amplitude of the wave vary from point to point.

Let us designate the amplitude of the wave function  $\Psi$  by  $\psi$ . For most problems, it is primarily  $\psi$  that interests us. If a vibratory process of frequency  $\omega$  occurs in a region, then, in the most general case:

$$\frac{\partial^2 \Psi}{\partial t^2} = -\omega^2 \Psi.$$

Therefore, a wave function will always satisfy the equation

$$\Delta \Psi = -\frac{\omega^2}{v^2} \Psi.$$

The part of the expression for  $\Psi$  that is a function of time always cancels in such an equation. Hence, the last equation is the equation for the wave *amplitude*  $\psi$ .

By means of the relation  $\lambda = v \frac{2\pi}{\omega}$  it may also be written in the form

$$\Delta\psi + \frac{4\pi^2}{\lambda^2} \psi = 0.$$

Sometimes this equation too is called the wave equation.

#### Sec. 129. PROPAGATION OF RADIO WAVES

Radio waves are propagated in accordance with the laws of reflection and refraction. In order to obtain the concrete results required for our discussion, it is merely necessary to generalise the theory to the case of a medium having a continuously varying coefficient of refraction. But this has already been done for elastic waves (see p. 109) and is completely applicable for electromagnetic waves, i.e., for light as well as radio waves. A wave travelling in a medium of variable  $n$ , i.e., a wave travelling with variable velocity, is propagated in such a manner that the least amount of time is taken to traverse the distance between two points. The path of the wave will be curvilinear and in passing from one layer of the medium to another layer where  $n$  is greater the wave will be deflected toward the normal to the boundary.

In order to determine the nature of the radio-wave propagation, the electrical properties of the Earth and the atmosphere must be known. The electromagnetic field of the wave is greatly affected by the magnitudes of the electrical conductivity and the dielectric constant of these two media.

How is the difference in behaviour of electromagnetic waves of different length explained? Of course, a significant role is played by dispersion. But an approximate indication of the behaviour of an electromagnetic wave may be obtained from an examination of the relationship between the displacement current and the conduction current. It is evident that a medium exhibits dielectric properties when the displacement current is much greater than the conduction current. On the other hand, if the displacement current is negligible, the medium may be considered to be a conductor.

The properties of the Earth's surface and the properties of the atmosphere must be examined from this standpoint.

Let us take a typical example. Experience in the field of radio engineering has shown that a flat terrain covered with trees may be characterised by a dielectric constant  $\epsilon$  of the order of 12 and a specific electric conductivity  $\gamma$  of  $7 \times 10^7$  (in the CGS system). To study the propagation of waves over the surface of a sea, it is important to know the values of  $\epsilon$  and  $\gamma$  for sea water. These values are 80 and  $10^{16}$ , respectively. The ratio of the conduction current density to the displacement current density (see p. 240 for the required formulas) is given by the formula

$$\frac{j_{cond}}{j_{displ}} = \frac{2}{c} \frac{\gamma \lambda}{\epsilon} = 0.7 \times 10^{-10} \frac{\gamma \lambda}{\epsilon}$$

in the CGS system. For long waves, say 2,000 metres, this ratio is equal to 77 for a wooded area and to 1,600 for a sea surface. The medium in each case, but especially in the latter, may be considered to be a good conductor. For short waves, say 20 metres, the first value decreases to 0.77 and the second to 16. This means that for short waves sea water continues to be basically a conducting medium, while a wooded area acts to a considerable extent as a dielectric.

Waves propagating over a conducting surface "cling" to this surface. The electric flux lines approach the Earth at right angles and travel along the terrestrial surface. That is why an electromagnetic wave can easily travel around the globe. (It takes

0.13 sec to do this and since this time can be measured quite accurately we can determine the propagation velocity of radio waves.) This applies to long waves. Short waves cling only to sea surfaces. In other regions, they behave like perfectly free waves. When such a wave travels along the surface of the globe, it penetrates the Earth and is absorbed. Moreover, the higher the oscillation frequencies, the greater the absorption.

A number of remarkable features in the behaviour of radio waves is explained by the presence in the upper layers of the atmosphere of a layer containing a large

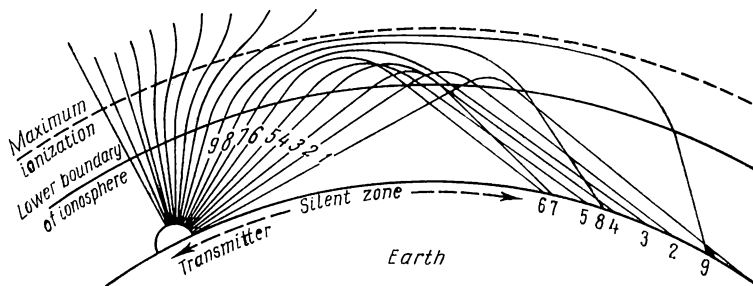


Fig. 139

number of free ions and electrons. This layer is known as the ionosphere. Thus, the region in which an electromagnetic wave travels may be roughly pictured as a dielectric bounded by two conducting layers.

Ionisation of the atmosphere is not uniform, i.e., the number of free charges per unit volume varies from one layer to the next. As was seen in Sec. 125, the coefficient of refraction decreases as the number of charges increases. Since the coefficient of refraction of a conducting medium is less than unity, a wave entering the ionosphere at an angle from a dielectric medium is deflected from the normal. The ionisation increases; hence, the deflection increases with each succeeding layer.

Furthermore, as Fig. 139 shows, a wave may either pass through the ionosphere and recede from the Earth or, after being bent more and more, return to the Earth. Roughly speaking—disregarding the nonuniformity of the ionosphere—a wave returns to the Earth if it strikes the ionosphere at an angle greater than the total internal reflection angle:

$$\sin \theta_0 > n = \sqrt{1 - \frac{4\pi N^2 e}{m\omega^2}}.$$

For smaller angles, the wave is propagated into outer space. By repeatedly being reflected from the ionosphere and the terrestrial surface, short waves can round the globe, experiencing considerably less energy losses than in the case of long waves.

Since UHF waves can pass through a layer of free charges, they are not reflected from the ionosphere. Therefore, radio reception on UHF is possible only along the line of sight.

The above picture of the atmosphere is greatly oversimplified. Investigations have shown that the density distribution of free electric charges in the atmosphere is characterised by several maxima since the ionosphere is composed of several layers. The stability of these layers differs, depending on the time of year. It is interesting that the existence of the layers is related to solar activity. Thus, variations in the state of the ionosphere corresponding to the 11-year sun-spot cycle

may be observed. Ionisation of the upper layers of the atmosphere is undoubtedly related to the arrival of cosmic radiation on the Earth.

From a study of the electrical properties of the ionosphere and the surface of the Earth, radio engineers have drawn a number of conclusions regarding the most favourable conditions for radio transmission and reception on waves of various lengths. However, we shall not go into this subject.

### Sec. 130. RADAR

A radar station consists of transmitting and receiving equipment. Every ten-thousandth of a second ( $\lambda$  in Fig. 140), the transmitter sends a pulse of duration  $\alpha$  (of the order of several microseconds) into space. If an object capable of reflecting the wave is intercepted in the solid angle "illuminated" by the radio waves, a part of the wave is reflected back to the radar station. The reflected signal is received  $\tau = \frac{2R}{c}$  sec after the pulse is transmitted into space.

This time may be measured by means of an oscilloscope. The sweep of the electron beam is synchronised with the transmitted pulses, and the demodulated signal from the receiver is fed to the second pair of oscilloscope plates. As a result, a "pip"

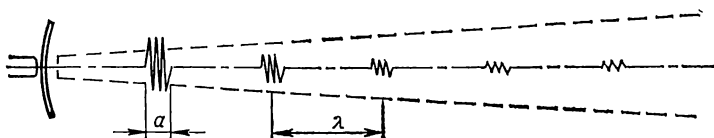


Fig. 140

displaced relative to the initial point of the sweep by a distance proportional to the time  $\tau$  appears on the oscilloscope screen. If the object intercepted by the radar pulse is stationary, the "pip" on the oscilloscope screen will also be stationary. Thus, synchronisation is achieved when the sweep time equals one ten-thousandth of a second, the interval of time between successive transmitted pulses. If the object "seen" by the radar is moving, then the pip on the oscilloscope screen also moves.

Modern radar systems are much more complex than indicated by this simplified picture. The motion of the electron beam of an oscilloscope from the centre to the edge of the screen is more complex than motion along a radius. While moving outward along the radius, the electron beam slowly rotates about the centre of the screen like the hand of a clock. This rotation is synchronised with the rotation of the radar antenna in such a manner that the illuminated line points in the same direction as the transmitted radio beam. In addition, the following important change in the operation of the oscilloscope is introduced: if the radio beam does not encounter an obstacle, and hence the receiver does not pick up a reflected signal, the oscilloscope screen remains dark. On the other hand, if a pulse is received, a spot on the screen is illuminated.

Thus, when a beam scans the horizon and encounters a body, this body is indicated on the oscilloscope screen by an illuminated spot. The distance of this spot from the centre of the screen is proportional to the distance of the radar from the object, and the azimuthal angle indicates the direction of the object.

Oscilloscope screens possess an afterglow; hence, an illuminated spot does not disappear while the radar is scanning the area, i.e., before returning to the same position. If the illuminated spot is due to the beam reflected from a fixed object,

the image on the oscilloscope screen is also fixed. If the object moves, a moving image will appear on the screen.

Due to the difference in the reflection coefficients of various objects, a characteristic picture of the region is depicted on the screen of a radar system with circular scanning. Rivers and lakes appear dark (little reflection), the Earth appears lighter and woods still lighter. Of course, metal objects are "seen" very clearly.

The nature of the visibility varies in accordance with the wavelengths used. Thus, for radio waves in the centimetre range, clouds are seen very clearly. Since longer waves are insensitive to clouds and rain, radar systems operating on such wavelengths can be used in all kinds of weather if they are not intended for the specific purpose of detecting clouds.

Radar principles find broad application in science and engineering. Thanks to radar, pilots experience no difficulty in conducting night flights and in landing on airports which are not illuminated. Radar is of great importance in meteorology. In addition to enabling us to detect rain and storm clouds at great distances or at night, which is essential for weather-forecasting, radar can be used to track

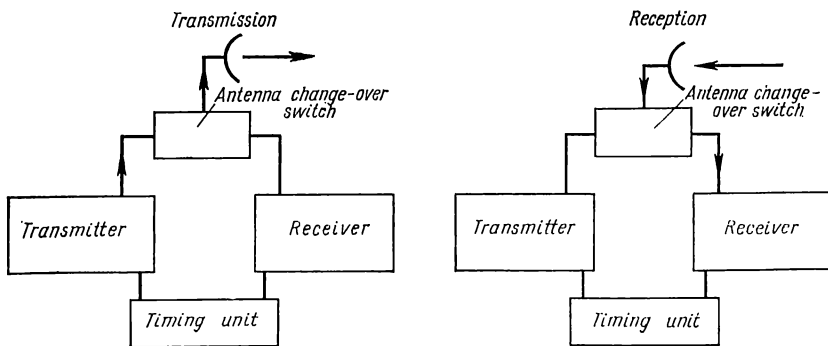


Fig. 141

meteorological balloons. Radar equipment installed on ships is a great aid to navigation safety, reducing to nil the possibility of accidental collisions of a vessel with other vessels or obstacles. In the field of astronomy, radar methods are used to determine the distance to meteors, and the direction and velocity of their flight. The waves are reflected, in the main, from the meteorite "trails", which consist of ionised gases. The Moon, the Sun and the planets are all within the reach of radar. Radar astronomy is of great practical importance. It has enabled us to develop navigational instruments to determine the location of a ship, from observations made on heavenly bodies, in any kind of weather and at any time of day or night.

A vast amount of the literature is devoted to radar problems. Since such problems rightly belong to the field of radio engineering and not physics, we have restricted ourselves to an elaboration of the principles of this remarkable development.

Fig. 141 shows a block diagram of a radar system.

# Interference Phenomena

## Sec. 131. ADDITION OF WAVES FROM TWO SOURCES

First, let us consider two ideal sources radiating spherical waves. Assume both sources are oscillating synchronously. In this case, for waves of any type, a characteristic field is created in which bright and dark "fringes" appear where the waves reinforce and cancel each other, respectively. This phenomenon is most easily demonstrated by means of water waves.

The mathematical calculations are straightforward. Take any point a distance  $r_1$  from one wave source and  $r_2$  from another. Then, maximum reinforcement of the waves occurs when the path difference  $r_1 - r_2$  equals a whole number of wavelengths,  $n\lambda$ . On the other hand, the waves annul each other when the path difference equals an odd number of half-wavelengths  $(2n + 1) \frac{\lambda}{2}$ .

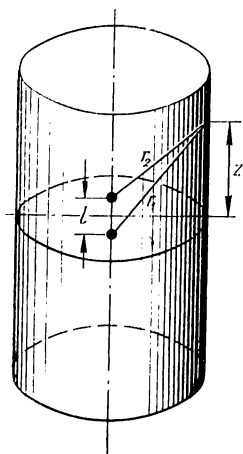


Fig. 142

We know from analytical geometry that a curved surface all of whose points satisfy the following condition is a hyperboloid: the difference between the distances to two foci is a constant. In Fig. 59 (p. 91) a plane is drawn through the wave sources. Shown in this plane are hyperbolas—loci for which the difference between the distances to the wave sources is a constant.

Now, let us consider the pattern obtained in a wave field on a cylindrical screen whose axis passes through the radiators as shown in Fig. 142. (We shall assume throughout that the radiators are point sources.) This interference pattern will consist of alternating bright and dark horizontal lines since the conditions are exactly the same for all points on the cylinder located at the same height, i.e., relative to the radiators. All such points are in the same state. A bright fringe will appear along a line around the middle of the cylinder; since the distance from both sources is the same, the waves reinforce each other. For points at a height  $z$  above the mid-line, the difference in the paths traversed by the rays,  $r_1 - r_2$ , will be represented by  $\frac{r_1^2 - r_2^2}{r_1 + r_2}$ . But  $r_1^2 - r_2^2 \approx 2lz$ , where  $l$  is the distance between the sources. Thus, the condition for the  $n$ -th bright fringe has the form

$$\frac{2lz}{r_1 + r_2} = n\lambda.$$

If the radiators are far from the screen, then for fringes close to the centre,

$$r_1 + r_2 \approx 2R,$$

where  $R$  is the radius of the cylinder. The bright fringes pass through the points  $z$  satisfying the condition

$$\frac{lz}{R} = n\lambda.$$

The distance between adjacent fringes is  $\Delta z = \frac{\lambda R}{l}$ .



*Example.* If two coherent sources (see Sec. 132) separated by a distance  $l = 1$  mm emit light of wavelength  $\lambda = 6,000 \text{ \AA}$ , then the distance between interference fringes on the surface of a cylinder of radius  $R = 1$  metre is  $\Delta z = \frac{\lambda R}{l} = 0.6 \text{ mm}$ .

If the light source emits waves of various wavelengths, the interference pattern will be coloured since the maximum conditions differ for different values of  $\lambda$ .

In addition to determining the positions of maximum and minimum interference, it is also of interest to determine the form of the intensity curve across the fringes.

Since  $\frac{lz}{R}$  is the path difference between the waves, then

$$\delta = \frac{2\pi}{\lambda} \frac{lz}{R}$$

is the phase difference, and the total amplitude at any point is given by

$$A \cos \omega t + A \cos (\omega t + \delta).$$

For equal amplitudes, we obtain the expression derived on p. 77:

$$2A \cos \frac{\delta}{2} \cos \left( \omega t + \frac{\delta}{2} \right).$$

The measured intensity (wave amplitude squared) is equal to the average value of this expression taken over the oscillation period.

Since

$$\left[ \cos^2 \left( \omega t + \frac{\delta}{2} \right) \right]_{av} = \frac{1}{2}$$

(see the following article regarding calculation of the average), then

$$I = 2A^2 \cos^2 \frac{\pi lz}{\lambda R}.$$

The intensity curve may be plotted as a function of the vertical coordinate  $z$  (Fig. 143).

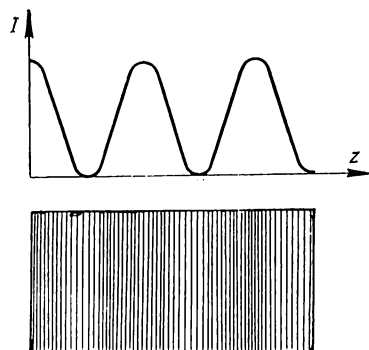


Fig. 143

## Sec. 132. COHERENCE

The superposition of two waves, described in the preceding section, may be physically achieved by various means and for various wavelengths. For example, two antennas radiating radio waves may be placed close together, two electric bulbs with point filaments may be placed close together, or an incident ray and the same ray reflected from a mirror may be brought together. Experiments show that by no means does interference occur in all cases. This phenomenon can best be investigated by examining the superposition of the fields of two antennas. It is easily shown that an interference pattern is obtained only when a phase difference that remains constant during the time of observation exists between the superimposed waves. In this case the oscillations are said to be *coherent*.

If the phase difference is fixed, the amplitude of the electromagnetic oscillations at a given point in space is constant. Thus, a maximum point remains a maximum and a point where the waves completely cancel each other always has zero intensity.

When the phase difference varies randomly, the pattern is completely different. During a certain interval of time the amplitude of the oscillations at a given point is a maximum, in the succeeding interval it assumes intermediate values, and then for an interval of time the waves cancel each other. If the duration of these intervals were commensurable with the practical capabilities of instruments, a fluctuating interference pattern could be detected. If the variations in phase difference are so rapid as to preclude detection by these instruments, the interference pattern is not revealed and the average value of the intensity is shown on the instruments. In such cases, we say that the oscillations are noncoherent.

What is the expression for the average intensity in a region where fields are superimposed? This is easily determined.

The amplitude of the total wave at a given point and at a given instant may be expressed in the form

$$A_1 \cos \omega t + A_2 \cos (\omega t + \delta).$$

The instantaneous intensity is proportional to this expression squared, i.e., it is equal to

$$A_1^2 \cos^2 \omega t + A_2^2 \cos^2 (\omega t + \delta) + 2A_1 A_2 \cos \omega t \cos (\omega t + \delta).$$

We are interested in the time-averaged intensity of the radiation, i.e.,

$$I = A_1^2 (\cos^2 \omega t)_{av} + A_2^2 [\cos^2 (\omega t + \delta)]_{av} + 2A_1 A_2 [\cos \omega t \times \cos (\omega t + \delta)]_{av}.$$

The average values of trigonometric quantities are encountered quite often in physics. It is therefore useful to recall that the average values of  $\sin x$  and  $\cos x$  are equal to zero, and the average values of  $\sin^2 x$  and  $\cos^2 x$  are equal to  $1/2$ , if the argument  $x$  of the trigonometric function assumes all values with equal probability. The average value of a function  $f(x)$  is by definition

$$[f(x)]_{av} = \frac{f(x_1) + f(x_2) + \dots + f(x_n)}{n}.$$

This formula can be used to calculate the average if the variable  $x$  assumes discrete values. But if the variable  $x$  is continuous and assumes all values in the interval from  $a$  to  $b$ , the formula for calculating the average value is obtained in the following manner. Divide the interval  $(b - a)$  into  $n$  segments of length  $\Delta x$ . Multiplying the numerator and denominator by  $\Delta x$ , we obtain

$$[f(x)]_{av} = \frac{f(x_1) \Delta x + f(x_2) \Delta x + \dots}{n \Delta x}$$

Going over to the limit, this takes the form

$$[f(x)]_{av} = \frac{1}{b-a} \int_a^b f(x) dx.$$

By means of this formula, we can calculate the average value of any function of a continuously varying random quantity. In calculating the average value of a periodic function, a single period should be used in determining the limits of integration, for the average value of one period is clearly equal to the average value of any number of periods. Thus, for example,

$$[\cos^2 x]_{av} = \frac{1}{\pi} \int_0^\pi \cos^2 x dx = \frac{1}{2}.$$

Let us write the formula for the intensity in the form

$$I = A_1^2 (\cos^2 \omega t)_{av} + A_2^2 [\cos^2 (\omega t + \delta)]_{av} + A_1 A_2 [\cos (2\omega t + \delta)]_{av} + A_1 A_2 (\cos \delta)_{av}.$$

Using our knowledge regarding the average values of  $\cos x$  and  $\cos^2 x$ , we obtain: for a phase difference between two waves varying randomly, i.e., for noncoherent

oscillations,

$$I = \frac{1}{2} (A_1^2 + A_2^2);$$

on the other hand, if the phase difference is fixed, i.e., if the oscillations are coherent,

$$I = \frac{1}{2} (A_1^2 + A_2^2 + A_1 A_2 \cos \delta)$$

or, for equal amplitudes,

$$I = 2A \cos^2 \frac{\delta}{2}.$$

This last formula is the same as the interference formula derived in the preceding article.

Radio waves radiated by neighbouring antennas may be made coherent or non-coherent by technical means.

As regards oscillations of light waves, one should first of all distinguish between the light of ordinary sources and that of lasers. At first glance it would seem that

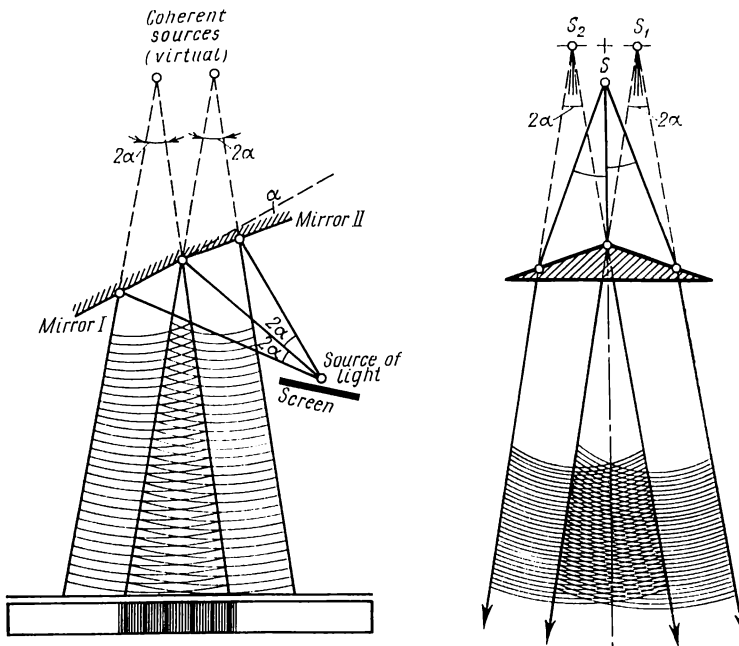


Fig. 144

it is impossible to achieve coherence since the radiations of individual atoms have no definite relationship to one another. The phases of the waves emitted by individual atoms are haphazardly distributed. It is quite natural that two light sources, no matter how closely they approximate point sources, do not yield an interference pattern. Nevertheless, coherent light oscillations do exist. They occur for rays "taken" from one and the same light wave.

Methods of artificially securing coherent sources are shown in Fig. 144. In one case two mirrors *I* and *II* that are slightly inclined relative to each other are used and in the other a double prism (biprism) is used to produce wave sources

from two virtual centres. Interference fringes may be observed on a screen placed anywhere in the interference field. The theoretical discussion presented above is entirely applicable to these cases; the nature of the fringes is determined by the distance between the virtual images of the light source and the distances from these images to the point of observation.

It is quite clear why the two parts of the "splitted" ray are coherent. Between any pair of atoms of a real source, there is no coherent relationship. However, by splitting such a ray into two parts, we enable the radiation of each atom to interfere with itself.

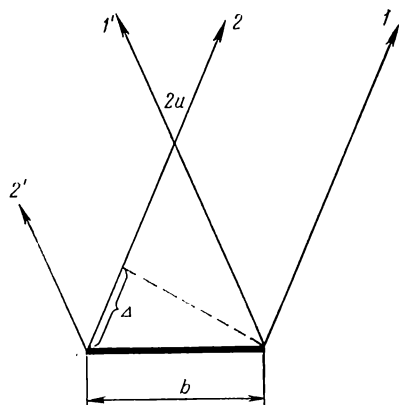


Fig. 145

Thus, consideration of various cases of light interference is reduced to the investigation of various cases of division of a light ray by reflection and refraction, and the successive superposition of the components of the splitted wave in the region of the interference field.

The size of the light source significantly affects the coherence of the "splitted" ray. Let us assume that the light source is of length  $b$  and that rays in the solid angle  $2u$  take part in the creation of the interference field (Fig. 145). Rays such as  $1$  and  $1'$  emanate from a single atom and are therefore coherent. The same is true of  $2$  and  $2'$ . Consider the interference due to the superposition of the fields

of rays  $1$ ,  $2$  and  $1'$ ,  $2'$ . For interference to occur, the fields of the coherent beams  $1$ ,  $1'$  and  $2$ ,  $2'$  must reinforce each other. The path difference  $\Delta = b \sin u$ , which exists between  $1'$  and  $2'$  as well as between  $1$  and  $2$ , tends to prevent this. Moreover, interference becomes possible only when  $b \sin u \leq \frac{\lambda}{2}$ .

Obviously, interference can occur only between light waves of equal wavelength. Consequently, the intensity of the interference fringe is determined not by the total radiated power, but by the power radiated by light waves of a given wavelength.

A limited coherent length of a light beam is a serious restriction for carrying out interference experiments with ordinary light. The point is that one radiation of the atom lasts during the time of the order of  $10^{-8}$  second. Taking into account the velocity of light propagation, it is not difficult to get convinced that the emitted wave train has a length of the order of a metre. Thus, the coherent length for a strictly monochromatic visible light will be of the order of a metre. If a high-pressure mercury lamp is used as a light source, then the coherent length will be of the order of only a millimetre. This means that the radiation of one and the same atom "fissioned" and brought together to one point with a path difference more than a millimetre (i.e. with a path difference exceeding the coherent length) yields no longer the phenomenon of interference.

Stimulated radiation of a laser whose origin will be considered in Sec. 168, is ideally monochromatic, i.e. the light frequency is millions of times higher than the frequency interval. The most important thing is that all atoms of the laser create a stimulated radiation on one phase. Consequently, in this case, radiations of different atoms are capable of interfering with one another. The total length of a wave train is of no importance since all laser atoms cause in-phase stimulated emission of radiation. Therefore, the coherent length loses its importance and the

"splitted" light of the laser will interfere even if one part of the beam covers a path tens and hundreds of metres longer than its other part. The "splitting" of a laser beam with the purpose of obtaining interference does not require the numerous measures which must be undertaken to ensure the observation of interference of ordinary light. It is sufficient to install the laser behind a screen with two slits and to create a common field of the light outgoing from these two slits.

Highly powerful laser light enables us to easily carry out interference experiments which were extremely difficult or even impossible before.

### Sec. 133. INTERFERENCE IN A PLATE

Let us investigate reflection and refraction of light incident on a flat plate of thickness  $d$  (Fig. 146).

Assume a plane wave impinges on the plate at an angle  $i$ . The beam of light is reflected and refracted. Moreover, the refracted beam strikes the lower surface of the plate and is also reflected and refracted. As a result, there arise numerous rays parallel to the primary reflected beam, and also numerous parallel rays transmitted in the second medium. All these rays are coherent and a phase difference exists between them. Hence, the conditions are present for interference in the reflected as well as in the transmitted rays.

As is well known, the reflection coefficient is not very large—at least, for normal incidence. In this case, the intensity of each "succeeding" ray is much less than the intensity of the preceding one. For example, for a reflection coefficient of 5%, the first reflected ray has an intensity of  $0.05 I_0$ . The second reflected ray undergoes two refractions and one reflection. Its intensity is  $0.95 \times 0.95 \times 0.05 I_0 = 0.045 I_0$ . Thus, the intensities of the first two rays are practically the same. But the third ray is much weaker since it undergoes three reflections and two refractions. Its intensity is equal to  $0.95 \times 0.95 \times 0.05 \times 0.05 \times 0.05 I_0$ , i.e., one *four-hundredth* of the intensity of the preceding ray.

Under conditions of small reflection coefficient, the phenomenon reduces to the observation of the interference of the first two rays.

As regards the transmitted rays, under conditions of small reflection coefficient the interference is not noticeable since the intensity of the second ray is one four-hundredth of the first, the intensity of the third is one four-hundredth of the second, etc. However, it is not very difficult to set up the experiment in such a manner that in the reflected as well as in the transmitted beam numerous interference rays occur.

If a monochromatic wave impinges on a flat plate, the interference pattern is determined by the phase difference between the first and second reflected rays.

From the wave formula

$$A \cos \omega \left( t - \frac{x}{v} \right),$$

it is evident that the phase of the wave, traversing the path  $x$  with velocity  $v$ , changes by  $\omega \frac{x}{v}$ , or  $\frac{2\pi}{\lambda} x$ , where  $\lambda$  is the wavelength in the medium. Designating

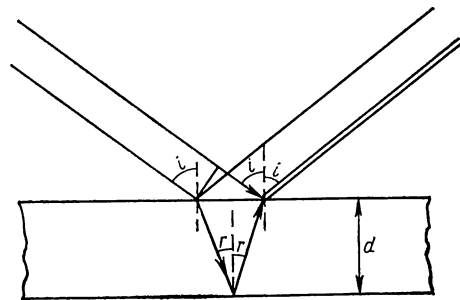


Fig. 146

the wavelength in vacuum by  $\lambda_0$  and recalling that the refractive index  $n$  is equal to  $\frac{\lambda_0}{\lambda}$ , the change in phase may be written in the form  $\frac{2\pi}{\lambda_0} nx$ . The product  $nx$  is often called the optical path of the wave. If a wave passes through several media, its phase changes by  $\frac{2\pi}{\lambda_0} S$ , where  $S = (n_1x_1 + n_2x_2 + \dots)$  is the optical path.

The phase difference  $\delta$  between the interfering waves, which determines the intensity of the resultant wave, is given by

$$\delta = \frac{2\pi}{\lambda} \Delta,$$

i.e., it is determined by the difference between  $S'$  and  $S''$ , the optical paths of these waves:  $\Delta = S' - S''$ .

Referring to Fig. 146, let us calculate  $\Delta$  for the case which interests us. It is most convenient to express  $\Delta$  in terms of the refraction angle  $r$ , the plate thickness  $d$  and the index of refraction  $n$ . As seen from the figure,

$$\Delta = 2dn \cos r.$$

However, in addition, it is necessary to take into account the phase jump occurring upon reflection (cf. p. 252). In this respect, the first and second rays differ, for the first is reflected from the external surface of the plate, while the second is reflected from the internal surface. Therefore, the electric vector of one of the rays undergoes a  $180^\circ$  phase jump and the other does not. Thus, the resultant phase difference is

$$\delta = \frac{2\pi}{\lambda_0} 2dn \cos r \pm \pi.$$

Maximum interference occurs when  $\delta = m2\pi$ , where  $m$  is a whole number; minimum interference occurs when  $\delta = m\pi$ . Therefore,

$$\text{maximum condition: } 2dn \cos r = m\lambda_0 \pm \frac{\lambda_0}{2};$$

$$\text{minimum condition: } 2dn \cos r = m\lambda_0.$$

Thus, depending on the value of  $\lambda$ ,  $n$ ,  $d$  and  $r$ , interference may cause the intensity of a wave reflected from a plate to be zero or a maximum. In an ideal experiment with a monochromatic beam, by varying the angle of incidence, for example, a reflected ray should alternately vanish and reappear. In an analogous experiment with a beam of white light, the plates should pass through all the colours of the rainbow in succession.

#### Sec. 134. FRINGES REPRESENTING EQUAL THICKNESS AND FRINGES REPRESENTING EQUAL INCLINATION

Several factors enter into the extremum condition  $2dn \cos r = m\lambda$ . Hence, if they are varied simultaneously, a confused picture may result. The effect is clearest when all the parameters, except one, may be considered fixed.

If a plate has a variable thickness  $d$ , a constant refractive index and a practically constant angle of incidence (and hence angle of refraction) for the portion of the plate under consideration, the interference will be observed in the form of *fringes representing equal thickness*. All parts of the plate having the same thickness  $d$  will be subject to the same conditions. Therefore, on an uneven plate, there appears a system of bright and dark fringes (or rainbow in case of white light). These lines connect points where the thickness of the plate is the same. This explains

the coloured fringes often seen on oily films spread on the surface of water. If a plate is wedge-shaped the fringes representing equal thickness consist of straight lines. Such fringes may be easily observed on soap films. In the case of a vertical film, the soap trickles down and the film becomes thinner in the upper region; horizontal fringes appear on the film.

When light impinges normally on a plate,  $\cos r \approx 1$  and fringes appear on the plate where the thickness  $d$  satisfies the relation  $2dn = m\lambda_0$ .

The difference in the thicknesses of the plate represented by adjacent fringes equals  $\frac{\lambda_0}{2n} = \frac{\lambda}{2}$ , i.e., a half-wavelength. Thus, bright fringes representing equal thickness indicate nonuniformities in plate thickness of the order of a tenth of a micron.

If the thickness from one point to another varies very slowly, the fringes may turn out to be very far apart. Thus, for example, in a dripping soap film a wedge may form having a 0.5-minute angle; in this case, as can be easily calculated with the aid of Fig. 147, the fringes will be 2 mm apart.

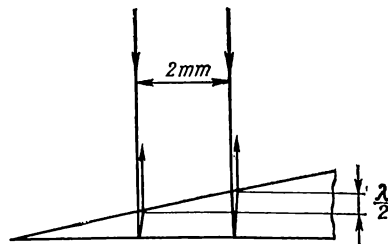


Fig. 147

If a wedge decreases to zero thickness, the end of the wedge appears dark in the reflected-light since thicknesses less than  $\frac{\lambda}{4}$  do not reflect light. The first bright fringe occurs for the thickness  $d = \frac{\lambda}{2}$  (the path difference is twice as great since the return path must also be included in the calculation). The next bright fringe occurs for  $d = \lambda$ , etc. Thus, the thickness may be determined by simply counting the fringes.

The question naturally arises: why are fringes representing equal thickness easily observed on thin films but not, for example, on a windowpane? The answer is that it is not possible to create the ideal conditions whereby the single variable quantity is the plate thickness  $d$ .

Let us consider the effect of a spread in the angle of incidence (refraction). If the angles vary from  $r_1$  to  $r_2$  in such a manner that on the interference maximum for  $r_1$  there is superimposed the extinction for  $r_2$ , the interference fringes will be smeared. What is the value of the angular interval  $\Delta r = r_2 - r_1$  that smears the pattern of fringes representing equal thickness? The value can be determined from the conditions:

$$2dn \cos r_1 = m\lambda \quad \text{and} \quad 2dn \cos r_2 = \left(m + \frac{1}{2}\right)\lambda,$$

whence

$$2dn (\cos r_2 - \cos r_1) = \frac{\lambda}{2}.$$

For simplicity, let us restrict ourselves to the case of normal incidence; let  $r_1$  equal zero and  $r_2$  a small value  $\Delta r$ . Then,

$$2dn = m\lambda \quad \text{and} \quad 2dn \left(1 - \frac{(\Delta r)^2}{2}\right) = \left(m + \frac{1}{2}\right)\lambda.$$

Hence,

$$2dn \frac{(\Delta r)^2}{2} = \frac{1}{2} \lambda$$

and since  $2dn = m\lambda$ ,

$$(\Delta r)^2 = \frac{1}{2m}.$$

If the plate is thin, the values of  $m$  are measured in units or tens of units. In this case, an angular spread of about a tenth of a radian, i.e.,  $5^\circ$ - $10^\circ$ , does not smear the pattern. However, for a plate 1 mm thick,  $m$  is already of the order of 5,000. Here, an angular spread of the order of only a hundredth of a radian suffices to prevent observation of fringes representing equal thickness.

But even if the geometry of the experiment is ideal, relatively thick plates do not yield an interference pattern. This is due to a limited coherent length.

In the case of laser light the difficulties of observing interference due to thick plates are mainly removed.

Now, let us consider fringes of another kind, namely, fringes representing *equal inclination*. Such fringes may be observed when a beam of light with a con-

tinuous spectrum of incident angles impinges on a plate having parallel surfaces, i.e., a plate for which  $d$  is the same at all points (Fig. 148).

Consider a beam of reflected rays contained in a given solid angle. Let us direct our attention to those rays lying along generating lines of a cone whose axis is normal to the plate. All rays lying on such a cone have the same value of  $r$  and yield lines representing equal inclination.

The differences in the method of observation of lines representing equal thickness and lines representing equal inclination should be noted. Since lines representing equal inclination

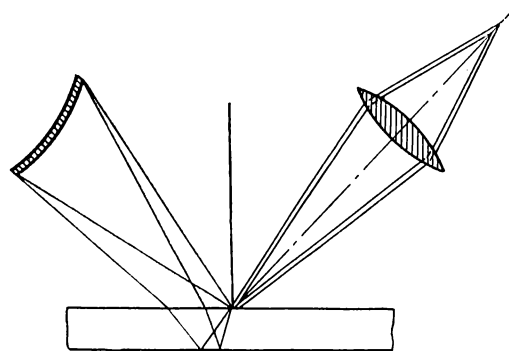


Fig. 148

occur at infinity, a lens must be placed in the path of the rays to make them visible. This enables us to observe curves representing equal inclination in the focal plane of the lens. For normal incidence, lines representing equal thickness may be observed with the naked eye on the surface of a wedge. If the light impinges on such a plate at an angle, lines representing equal thickness may be observed on the surface of a wedge only in the case of very thin films. Otherwise, the interference pattern is observed in two planes located above and below the wedge at a distance  $d \frac{\sin i}{\sin \alpha}$ , where  $\alpha$  is the wedge angle. To derive this formula—which we leave to the reader—plot a ray incident on the surface of the wedge at an angle  $i$  and the rays reflected from the upper and lower surfaces. The observation plane of the interference pattern will pass through the point where the extensions of the two reflected rays intersect. Note that the above formula is only valid for the case of an air wedge.

#### Sec. 135. PRACTICAL APPLICATIONS OF INTERFERENCE

Interference methods are widely used for the measurement of small distances and small changes in distances. They enable us to detect thickness changes of less than one-hundredth the wavelength of light. An accuracy of  $10^{-7}$  cm may



be achieved in the measurement of the unevenness of a crystal surface by interference methods.

Many applications are based on the use of curves representing equal thickness. This method is widely used in the optical industry. For example, in order to check the quality of the surface of a glass plate, an air wedge is created between the plate under test and a standard plate having an ideal flat surface and the fringes representing equal thickness are examined. The air wedge is formed by pressing the two plates together along one edge. If both surfaces are flat, the lines representing equal thickness will be parallel straight lines.

Let us assume that the surface of the plate under test has a depression or a bump. The lines representing equal thickness will then be distorted, i.e., they will bend

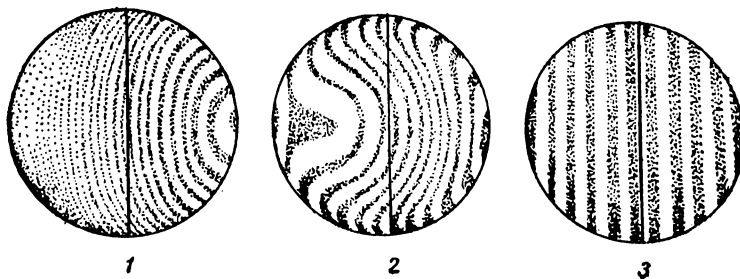


Fig. 149

around the defective area. If the angle of incidence of the light is varied, the fringes are displaced in one direction or the other, depending on whether the defect is a depression or a bump. The patterns seen under a microscope in these cases are shown in Fig. 149. The first two pictures are those of defective samples. In the first the defect is located on the far right, while in the second the defect is on the left. The third picture is that of a sample without defects.

This method may also be used for very accurate measurement of the coefficient of expansion. For this purpose, an air gap must be created between the surface of the object under test and a perfectly flat surface. As the object expands, the thickness of the air layer changes and the fringes representing equal thickness begin to move. If a line is displaced to such an extent that the next one takes its place, the thickness of the air layer at this location has changed by  $\frac{\lambda}{2}$ . If, as is usually the case, the measurement is performed using monochromatic light, the fringes are very sharply delineated and the displacement of a line by one-hundredth of the distance between lines may be measured.

Accurate measurement of the refractive index of a substance may be performed by means of an interference refractometer. In such an instrument, the interference between two light rays that are separated as much as possible is observed (Fig. 150). For this purpose, a thick plate is used and a convenient angle of incidence selected (for ordinary glass, the best angle is about  $50^\circ$ ). The rays travelling between the plates are separated and the substance being tested is placed in the path of one of them. This changes the optical path of this ray and hence the path difference between the interfering rays. If the interferometer plates are exactly alike and perfectly parallel, the interfering rays cover the same distance and reinforce each other. If the plates are inclined with respect to each other, a path difference is created and the clarity of the field being observed is reduced.

Such is the situation for a beam of perfectly parallel rays. But if a slightly divergent beam impinges on the plate, a system of fringes representing equal inclination is seen through the eyepiece. In this case, the variations in the optical path difference are conveniently determined by counting the interference fringes passing the cross hair of the instrument.

Let us assume that a body of length  $l$  and refractive index  $x$  is placed in the path of one of the rays. If the refractive index of the medium is  $n_0$ , the optical path difference changes by  $\Delta = l(n - n_0)$ . Therefore,  $\frac{\Delta}{\lambda}$  fringes should pass through the eyepiece of the instrument. The accuracy of this method may be easily gauged from the fact that a displacement of one-tenth of the distance between lines is

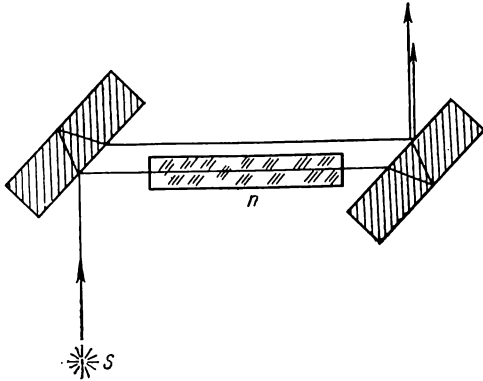


Fig. 150

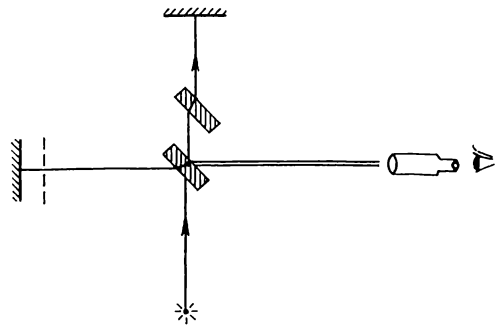


Fig. 151

easily detected. For such a displacement,  $\Delta = 0.1\lambda = 0.5 \times 10^{-5}$  cm, which for a length  $l = 10$  cm enables us to distinguish a change of  $0.5 \times 10^{-6}$  in the refractive index.

The well-known Michelson interferometer (Fig. 151) is used for the accurate measurement of length as well as for the determination of the velocity of light (see p. 316). In this instrument, a parallel beam of monochromatic light impinges on a glass plate having parallel surfaces, one of which is covered with a translucent layer of silver. This plate is placed at a  $45^\circ$  angle to the incident beam. As a result, the beam is divided into two parts. One part moves parallel to the extension of the incident beam and the other is directed perpendicular to the incident beam (to the left). These rays impinge normally on two mirrors and return to the same points on the translucent plate from which they came. Each ray returning from the mirror is repeatedly divided at the plate. Part of the light returns to the source and the other part enters the telescope to the right. As a result, two coherent interfering rays appear in the field of the telescope. It is seen from the figure that after the first division at the lightly silvered surface the ray coming from the mirror opposite the telescope passes through the half-silvered plate twice. Therefore, in order to provide equal optical paths, the ray coming from the other mirror is passed through an equalizing plate identical to the first plate, but without the translucent layer of silver.

In the field of the telescope there appear lines representing equal inclination (rings), corresponding to interference in an air plate whose thickness is the difference in the distances of the mirrors from the translucent layer. The displacement of one of the mirrors by a quarter of a wavelength corresponds to the transition from a maximum to a minimum, i.e., it results in the displacement of the pattern

by "half a ring". Such a change may be easily detected by an observer. Thus, the sensitivity of an interferometer using rays of violet light is better than  $1,000 \text{ \AA}$ , i.e.,  $0.1 \text{ micron}$ .

An interesting application of the Michelson interferometer principle is the microinterferometer developed by the Soviet physicist V. P. Linnik. In this instrument, one of the interferometer mirrors is replaced by an object to be investigated. Lines representing equal thickness are observed in the image plane. Dimension details of  $0.1 \text{ micron}$  yield a sharp transition from maximum to minimum illumin-

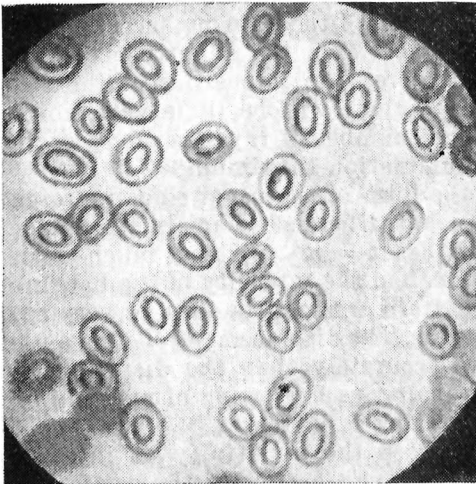


Fig. 152

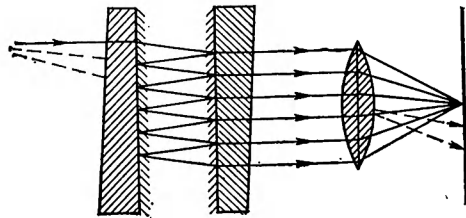


Fig. 153

ation. A microinterferometer is generally constructed in the form of an adapter to a usual microscope and screws into the drawtube in place of the objective. Of great importance in science are interference microscopes, which in principle are similar to the microinterferometer. (In the microscope, interference does not occur in the image plane, but occurs rather in front of the usual objective, i.e., the interference image has the same dimensions as the microobject). A double-ray interference microscope does not yield a gain in magnification. The advantage of this method lies elsewhere. For example, we often encounter objects of microscopic investigation that are either entirely transparent or vary little in transparency from one point to another. Before this method was developed, the details of such objects could be made visible only by dyeing (generally speaking, different structural elements absorb dye differently). But dyeing has little application in the investigation of living microorganisms. Fig. 152, which shows a microphotograph of frog's blood magnified 300 times, illustrates the possibilities of the interference microscope. The drawbacks of this method include considerable loss of light in the interference mechanism and complexity of the microscope's optical system. A further increase in the sensitivity of interference methods and, therefore, a further advance in the field of interference microscopy is possible by going over from a double-ray interferometer, to a multi-ray interferometer. In double-ray interferometers, the screen illumination is proportional to  $1 + \cos kh$ , where  $h$  is the displacement along the screen. Due to the smoothness of the transition from maximum to minimum illumination, a small displacement of the interference fringes

is difficult to detect. In multi-ray interferometers, the situation is considerably improved. As an example, let us consider the Fabry and Perot interferometer. A Fabry and Perot interferometer (Fig. 153) consists of two rather thick glass or quartz plates partially silvered on their adjacent sides. These surfaces are perfectly parallel to each other and the region between the plates contains air. When a beam of light enters the interferometer, the beam is divided into a transmitted and a reflected components every time it impinges on one of the layers of silver. Thus, in the transmitted as well as in the reflected light, there is obtained a set of coherent light beams whose intensities decrease in geometric progression (see p. 268) and whose phases are displaced in arithmetic progression. In order to secure interference of a large number of rays, the decrease in amplitude during successive reflections must not be large. This condition is met when the silver coating on the plates has a reflection coefficient of 0.9 or more. The intensities of the rays in the transmitted light are then quite low, but there will be little variation from ray to ray. This makes it possible for a large number of rays (as many as 10-15) to take part in the formation of each illumination maximum.

The interference pattern is obtained in the form of the usual rings representing equal inclination, but with one very important difference, namely, the principal maxima determined by the condition  $2dn \cos \alpha = m\lambda$  are now much narrower and their intensities are tens of times greater than the intensity of the background between them. Therefore, the interference pattern assumes the form of very narrow bright fringes separated by broad dark intervals. The displacement of such a narrow maximum may be determined much more accurately than the displacement of a fringe in a double-ray interferometer. An analogous narrowing of maxima occurs when the number of slits in a diffraction grating is increased.

Thus, multi-ray instruments sharply increase the sensitivity of interference methods. Such systems are indispensable in the investigation of the vertical structure of an object surface in reflected light. The magnification of details in the vertical direction may reach a value of 400,000, which makes it possible to reliably resolve details of the order of 5-10 Å. This is only several times greater than the distance between atoms! An example of such photography is the picture of the spiral growth of a crystal shown on p. 508.

At the present time, interferometers operating on natural light sources are replaced by those using lasers. The merits of the latter are quite obvious: incomparably high power of light, strict monochromaticity, ideal parallelism of the light pencil and unlimited coherent length.

With the aid of lasers astronomers can carry out measurements on a 200-inch telescope with an interferometer whose one arm has a length of a dozen of metres, the other being several centimetres long.

Interferometers used to control lens sphericity can be manufactured with only one comparing surface, whereas, using ordinary light, with a change in radius of lenses under trial, the operator had to change the comparison pattern as well, to say nothing of the fact that interferential pictures have become incomparably brighter, and therefore are analyzed readily and more precisely.

The possibility to manage without compensating the optical path of one of the rays enables the manufacturer to produce interferometers of an entirely new type. It has become possible to follow displacements of dams, geological drift, as well as crust tremors.

Of course, it was known before that, due to Doppler effect, it is possible to measure the rate of displacement of one of the interference mirrors, by measuring the rate of motion of interference rings.

# Scattering

## Sec. 136. SECONDARY RADIATION

Under the action of an electromagnetic wave, every molecule becomes a secondary radiator of electromagnetic waves. Due to the electric force, the electron cloud is displaced relative to the atomic nuclei and the molecule acquires a dipole moment varying in time as the frequency of the incident wave. The behaviour of such a molecule differs in no way from the behaviour of the elementary dipole discussed in Chapter XX. The intensity of the secondary wave is given by the formula derived on p. 246 (intensity  $\sim \frac{\omega^4}{R^2} \sin^2 \theta$ ) and the spatial intensity distribution of the secondary radiation is shown in Fig. 133.

In a number of cases, which will be discussed below, the phenomenon of secondary radiation leads to various phenomena of electromagnetic wave scattering. By scattering, incidentally, we generally mean any electromagnetic wave propagation phenomenon that is not included under refraction, reflection and rectilinear propagation.

The intensity formula given above is valid for any electromagnetic wave. However, the fact that the intensity increases sharply with radiation frequency explains why the effects of wave scattering by a molecule are not detectable when the wavelengths are very long. The scattering intensity of visible light is quite sufficient to produce significant effects.

Light wavelengths are hundreds and thousands of times greater than the dimensions of ordinary molecules. Therefore, all the electrons of a molecule are made to vibrate in the same phase by the external field. For light waves, ultraviolet rays and even very soft X-rays (i.e., X-rays of long wavelength), a molecule behaves like an elementary electric dipole.

The picture changes considerably in the case of X-rays having a wavelength of the order of  $1 \text{ \AA}$ . Now, the dimensions of the molecule are larger than a wavelength and different portions of the molecule's electron cloud vibrate in different phases. In order to determine the intensity of the scattered wave, we must take into account interference effects occurring between waves scattered by different parts of the molecule.

In principle, this is not very difficult. First, it is necessary to divide the molecule's electron cloud into small volumes. The dimensions of each such volume,  $\Delta v_h$ , must be much less than a wavelength. Then, the electrons in this volume will scatter in the same phase. Designating the density of the electron cloud by  $\rho$ , we obtain  $\rho \Delta v_h$  electrons in a volume  $\Delta v_h$ . The amplitude of the secondary wave created by the  $h$ -th volume is proportional to  $\rho \Delta v_h$ . The amplitudes of the scattered waves are added with due regard to the phase differences between the elementary waves and this sum is then squared. It is found that the intensity distribution of the scattered wave differs significantly from the radiation pattern of a single dipole. This is understandable, of course, for there will be directions in which elementary waves scattered by different volume elements reinforce each other, i.e., act in phase, and, on the other hand, directions in which elementary waves tend to annul each other. An important conclusion to be drawn from such calculations is the following: when interference occurs between elementary waves emanating

from different volume elements of a particle, waves travelling backwards tend to annul each other in the final analysis; on the other hand, those travelling forward reinforce each other.

We have been discussing secondary radiation from a molecule, but often the secondary radiator is a much bigger particle, namely, one composed of numerous molecules. It may be a particle of dust, colloidal substance, fog, crystalline substance, smoke, or a large albuminous molecule, etc. The nature of wave scattering by particles is determined by the ratio of their dimensions to the wavelengths of the exciting electromagnetic wave. If the particle is small relative to the wavelength, the wave is scattered as by a single elementary dipole. If this is not the case, interference effects occur and the forward scattering predominates.

Different parts of a particle may possess different scattering powers. This is precisely the situation in the case of X-rays scattered by a molecule. A particle whose scattering power is the same throughout the volume is the simplest body from the standpoint of scattering investigations. We shall confine our attention to such a system, for not only are the calculations simple in this case, but, in addition, a system of this kind is easily simulated experimentally by an aperture in an opaque screen.

#### Sec. 137. WAVE DIFFRACTION AT APERTURES

The amplitude of a wave scattered by a particle is determined by the distribution of scattering material in the particle. Particles ("apertures") may be encountered in which the density of the scattering material gradually decreases with increasing distance from the centre of the atom. On the other hand, more pronounced nonuniformities may be encountered, e.g., inclusions and pores at whose edges the density changes abruptly.

Peculiar diffraction effects arise when scattering takes place on such nonuniformities. With increasing angle the scattering intensity first gradually decreases to zero. Then, as the angle increases further, the intensity increases to a maximum value, whereupon it again decreases to zero. Subsequently, the wave-like nature of the curve continues with decreasing amplitude. Scattering on such objects leads to the formation of diffraction bands and spots of various shape depending on the nature of the scattering object.

The most pronounced diffraction effects are observed for scattering at apertures made in an opaque screen. Every aperture may be viewed as a region uniformly filled with radiating dipoles. The scattering patterns of an aperture and a particle having the shape of such an aperture should yield identical curves of intensity vs. scattering angle.

For light rays, diffraction patterns are best observed using parallel rays in accordance with the following scheme. The rays of a light beam emanating from a source are made parallel and allowed to impinge on a screen in which various inclusions (if the screen is transparent) or apertures (if the screen is opaque) are located. A lens placed behind the screen brings the parallel rays into the plane of photographic plate or screen for the observation of the effect. If there are no nonuniformities, apertures, etc., in the path of the rays, the lens gathers the rays in a point. Otherwise, a scattering or diffraction pattern appears on the screen.

Fig. 154 shows the diffraction patterns obtained in this manner from (a) two needles and a thin wire and (b) a circular aperture. To make these patterns more

understandable, let us determine the intensity distribution of scattered radiation\* for the simple case of an aperture having the shape of a slit.

Let a wave impinge normally on a slit cut in an opaque screen. Considering the slit divided into volume elements  $\Delta V$  as shown in Fig. 155, let us write the expression for the wave emanating from an arbitrary volume element  $\Delta V_h$  at an angle  $\varphi$  to the incident wave. The waves from different volume elements  $\Delta V_h$  arrive at the point of observation in different phases. If the path difference is

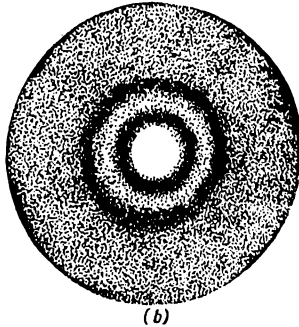
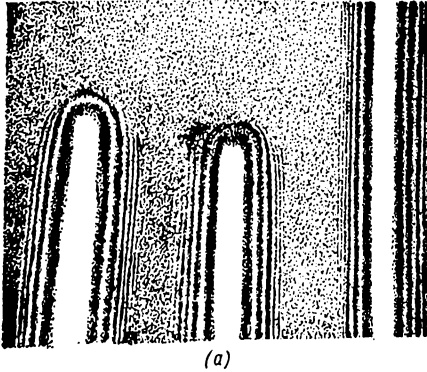


Fig. 154

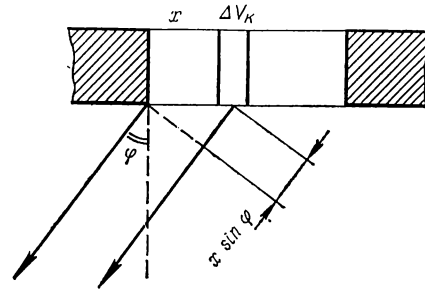


Fig. 155

measured relative to the outermost ray (on the side of deflection), the rays emanating from the other volume elements will cover a distance that is greater by the amount  $x \sin \varphi$ . Hence, these rays will be displaced in phase by  $\frac{2\pi}{\lambda} x_h \sin \varphi$ .

The amplitude of the wave scattered by the  $k$ -th volume element is proportional to the "scattering" volume element  $\Delta V$ , i.e., to the expression

$$\Delta V_h \cos \left( \omega t - \frac{2\pi}{\lambda} x_h \sin \varphi \right).$$

It is necessary to take the summation of the expressions for all volume elements. Instead of taking the summation we may integrate with respect to  $x$ , where the  $x$ -axis is taken along the width of the slit. Replacing  $\Delta V_h$  by  $\Delta x_h$ , which is proportional to it, and going over to the limit, we obtain for the amplitude of the

\* "Scattered radiation" and "diffracted radiation" have exactly the same physical meaning. The terms "diffraction" and "diffracted" are generally used when the scattering pattern has rather pronounced maxima and minima. When the nature of the interference pattern is not so evident, we speak of "scattering".

scattered wave at the angle  $\varphi$ :

$$A = k \int_0^a \cos \left( \omega t - \frac{2\pi}{\lambda} x \sin \varphi \right) dx,$$

where  $k$  is a coefficient of proportionality and  $a$  is the width of the slit.

Introducing the variable

$$z = \omega t - \frac{2\pi}{\lambda} x \sin \varphi,$$

we obtain

$$dz = -\frac{2\pi}{\lambda} \sin \varphi dx$$

and, therefore,

$$A = \frac{k}{\frac{2\pi}{\lambda} \sin \varphi} \left[ \sin \omega t - \sin \left( \omega t - \frac{2\pi a}{\lambda} \sin \varphi \right) \right].$$

Designating  $\frac{\pi a}{\lambda} \sin \varphi$  by  $u$  and performing a trigonometric transformation, we obtain

$$A = \frac{ka}{u} \sin u \cos (\omega t - u).$$

Thus, the resulting oscillation at the point of observation has the amplitude  $\frac{ka}{u} \sin u$ , i.e., the observed intensity is

$$I = k^2 a^2 \frac{\sin^2 u}{u^2}.$$

This is the formula for the intensity distribution as a function of the scattering angle.

In most diffraction experiments, we are interested in small values of the scattering angle  $\varphi$ . The reasons for this will become clear later. Therefore, replacing  $\sin \varphi$  by  $\tan \varphi$  and since

$$\tan \varphi = \frac{x}{f},$$

where  $x$  is the distance of the point of observation in the plane of the photographic plate to the centre of the diffraction pattern and  $f$  is the distance from the slit to the plate, we obtain for  $u$  the expression

$$u = \frac{\pi a}{\lambda} \frac{x}{f}$$

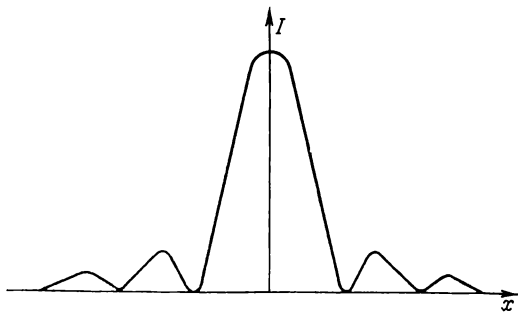


Fig. 156

Figure 156 shows the  $\frac{\sin^2 u}{u^2}$  curve.

Since  $u$  is proportional to  $x$ , this corresponds to the diffraction pattern appearing on the plate.

The locations of the dark fringes are easily determined from the condition  $u = \pm n\pi$ , where  $n$  is a whole number. Thus, the first zero occurs at

$$x = \frac{\lambda f}{a}$$

This value of  $x$  also represents the distance between two successive positions of zero intensity



From this formula, we can determine when diffraction phenomena may be detected for various wavelengths under various conditions.

The diffraction of light ( $\lambda = 0.5 \times 10^{-4}$  cm) may be clearly observed under laboratory conditions if the aperture is of the order of 0.1 cm and the distance between the screen and plate is of the order of 2 metres. Since for these values  $x = 1$  mm, the effect is clearly visible.

Rays of visible light yield noticeable diffraction from a tennis ball ( $a = 5$  cm), but at a greater distance. If the distance  $f$  is equal to 100 metres and the wavelength  $\lambda$  is 5,000 Å, then  $x = 1$  mm. Thus, in this case too the distance between positions of zero intensity of the scattered radiation is of the order of magnitude of 1 mm.

Diffraction of radio waves may also be observed if a proper choice of conditions, as determined by the equation  $x = \frac{\lambda f}{a}$ , is made.

Assume that the values of  $f$  and  $\lambda$  are fixed. The width of the slit greatly affects the diffraction pattern. If the width of the slit is large,  $x \rightarrow 0$ , i.e., the slit image focussed by the lens is infinitely narrow. As the width of the slit is decreased, the diffraction pattern begins to take shape and the first diffraction minimum begins to move further and further away from the centre of the pattern. Finally, when the slit is made so narrow that our approximation in the formula for  $u$  (the substitution of  $\sin \varphi$  by  $\tan \varphi$ ) is no longer justified, the image of the slit on the screen becomes smeared. A still further decrease in width, to the point where the wavelength and the width of the slit become equal, results in the slit yielding secondary radiation as a single source. The interference of primary waves disappears and the primary wave is radiated from the slit in all directions.

For apertures and particles (or inclusions) having other shapes, the diffraction patterns appear entirely different (see Fig. 154). Nevertheless, the general laws remain valid and the basic features of the pattern are maintained. Thus, for example, in the case of diffraction from a circular aperture or other circular nonuniformity, concentric rings are formed and the diameter of the smallest dark ring is  $1.22 \frac{\lambda f}{D}$ , where  $D$  is the diameter of the aperture.

Since diffraction patterns have maxima at different locations for different wavelengths, white light is resolved into its spectrum upon diffraction. Therefore, diffraction from a circular particle or aperture has the form of a rainbowed ring.

## Sec. 138. A SYSTEM OF RANDOMLY DISTRIBUTED SCATTERERS

We have considered the behaviour of various secondary radiators of electromagnetic waves as a function of the ratio of their dimensions to the wavelength of the incident wave. But the properties of a radiator are only roughly determined by its dimensions. The detailed pattern is determined by the distribution of matter in the scattering particle. Only when the dimensions of a particle are small relative to the wavelength is the distribution of matter in the particle of no consequence. In this case, the particle scatters as a whole, i.e., as a single electric dipole. When this condition is not satisfied, the pattern is complex since it is determined by the interference of waves scattered from various volume elements of the particle. We have considered only one example of a scatterer whose dimensions are larger than a wavelength, namely, a homogeneous scattering particle, which may take the form of an aperture in an opaque screen.

Now, let us consider the problem of scattering by systems of particles—e.g., a system of gas molecules, specks of dust or smoke particles; a system of hoarfrost crystals on a windowpane; or a system of holes in a piece of gauze. In all

such cases, the pattern becomes complex due to the fact that the electromagnetic waves emanating from the various scatterers may, generally speaking, interfere with each other. Now, the scattering pattern will depend not only on the properties of a scattering particle, but on the arrangement of the particles. Thus, it is important to know how close the scatterers are to one another and whether their spacing is regular or random. Then, depending on the circumstances, the waves scattered by the various particles may interfere to a maximum extent, partially or not at all.

Confining ourselves to the extreme cases, let us first consider scattering by a system of randomly distributed particles—e.g., the scattering of X-rays by a large cluster of atoms or molecules having a random distribution.

In the case of a large number of identical scattering centres (e.g., atoms, molecules or larger identical particles), the resulting scattering is determined, as we have already stated, by the scattering of a single centre (region) and the arrangement of the scattering centres. The scattering pattern in the case of uniform distribution of scattering centres is quite different from the pattern in the case of random distribution.

If the scattering centres are randomly distributed as, for example, in the case of gas molecules, the waves scattered by different centres may be considered to be noncoherent. This is because in the case of randomly distributed scattering centres the phase relationship between waves emanating from different centres is quite arbitrary. It is safe to say that the number of waves with positive amplitudes arriving at a point of observation (from different centres) will be exactly equal to the number of waves with negative amplitudes arriving there. The result is clear. Designating the amplitudes of the waves from the various centres by  $A_1$ ,  $A_2$ ,  $A_3$ , etc., the total amplitude at the point of observation is

$$A = A_1 + A_2 + A_3 \dots$$

Since the intensity is proportional to the amplitude squared, we obtain

$$I \sim A^2 = A_1^2 + A_2^2 + A_3^2 + \dots + 2A_1A_2 + 2A_1A_3 + \dots + 2A_2A_3 + \dots$$

But among the double products there will be just as many positive terms as negative terms. Therefore, with a high degree of accuracy, the total is given by the sum of the amplitude-squared terms. In other words, the total intensity scattered by identical randomly distributed centres is expressed as follows:

$$I = NA^2,$$

where  $N$  is the number of scattering centres and  $A$  is the scattering amplitude of one centre.

Thus, it turns out that scattering by a large number of randomly distributed particles is very similar to scattering by a single particle. The only difference is that the effect is  $N$  times greater.

Data on scattering by a single molecule are obtained by investigating the scattering of X-rays by a gas. Fig. 157 illustrates a laboratory set-up for the study of X-ray scattering by gases. A beam of X-rays is made monochromatic by reflecting it from a crystal, whereupon it is directed into a gas chamber. Scattering is determined by its action on a photographic film, i.e., the degree of blackening is a measure of the scattering intensity. We can thus obtain the intensity as a function of scattering angle—rapidly decreasing smooth curves for monatomic gases and decreasing curves having small maxima for polyatomic gases. By means of these curves and theoretical formulas, the electron-density distribution in a molecule may be determined.

There are many examples of systems that scatter electromagnetic waves as a gas scatters X-rays.

The scattering of light rays in a room with dust is a familiar example. Through a chink in a window curtain, a narrow, rectilinear beam of light visible from all sides penetrates a room. Light waves act on a system of dust particles in much the same way as X-rays act on a system of molecules. The distances between specks of dust are quite large and the distribution of the particles is completely random. Hence, there is no interference between the waves scattered by different dust particles, and the scattering pattern is similar to that created by a single speck of dust. The sole difference lies in greater intensity, i.e., the intensity is proportional to the number of dust particles in the field of the primary beam of light. Each speck of dust behaves as an elementary electric dipole, for the dimensions of such a particle are less than the wavelength of light. Therefore, the laws applicable to the scattering of light by dust particles, i.e., the dependence on the wavelength of light and the nature of angular distribution, are the same as for an elementary electric dipole (i.e., the intensity formula given on p. 245 and the intensity distribution shown in Fig. 133 are applicable here too).

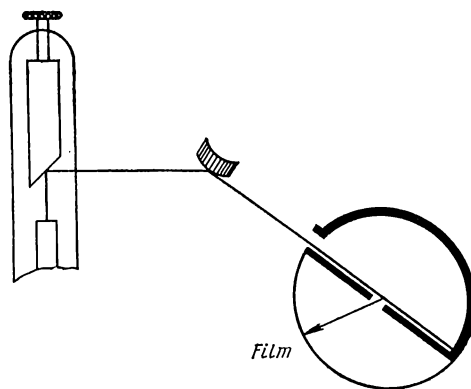


Fig. 157

This agreement can also be easily demonstrated by comparing the diffraction pattern from a single aperture with that from a system of randomly distributed apertures. Experiments show that, as regards the relative intensity distribution of scattered light, these two diffraction patterns are entirely alike. Of course, in the case of  $N$  apertures, the intensity of the light scattered by the screen is  $N$  times greater than the intensity of the light scattered by an opaque screen having one aperture.

Since the scattering pattern due to a large number of randomly distributed centres has the same form as that due to a single centre, it becomes clear why a lantern observed through a window covered with hoar-frost has the familiar rainbowed halo. This pattern is simply the result of diffraction from the ice particles. Since their distribution is perfectly random, they behave as "circular" particles.

#### Sec. 139. BEHAVIOUR OF A PERFECTLY HOMOGENEOUS MEDIUM

In this article we shall consider the other extreme case of the phenomenon of scattering. By a perfectly homogeneous medium we mean a system of scatterers that are distributed uniformly and continuously throughout a given region. Transparent glass constitutes such a medium with respect to light waves. Since the wavelength of light is considerably greater than the distance between the atoms in glass, a piece of transparent glass may be considered to be divided into volume elements that are considerably smaller than the wavelength, but at the same time still containing a large number of molecules. The medium may be considered homogeneous if the number of molecules in all such volume elements are approximately the same.

The scattering of electromagnetic waves is not only a function of uniformity in regard to the number of radiators per unit volume but is also a function of uniformity in the orientation distribution of the radiators. In the final analysis, the

scattering power of a body is determined by its dipole moment, which is composed of the individual dipole moments of the molecules contained in this volume. It may be stated, therefore, that the scattering power of a body is determined by the value of the dielectric constant or, since  $\epsilon = n^2$ , by its refractive index. Hence, for a given wavelength, a medium that is uniform as regards the scattering of electromagnetic waves must also be uniform as regards the refractive index.

Observations on the behaviour of electromagnetic waves in perfectly homogeneous media show that no scattering takes place in such media. Thus, when a ray of light passes through a transparent body, the ray cannot be seen from the side (compare this with the case of a light ray entering a room with dust).

Each volume element of a homogeneous body is a wavelet source. Nevertheless, wave scattering does not occur. Only one explanation is possible, namely, the wavelets scattered by a homogeneous medium in any direction at an angle to the primary ray are completely annulled due to interference. This theorem is capable of rigorous proof but the proof will be dispensed with since it is quite evident that this is the only possible explanation.

Nevertheless, the phenomenon of scattering does make itself felt quite considerably in a homogeneous medium. Scattered waves annul each other in all directions except one, namely, the direction in which the primary wave is propagated. Forward scattering does not represent simply a superposition on the primary wave but a change in its velocity as well.

It turns out that the phenomenon of electromagnetic wave refraction, which we have already discussed, may be interpreted as a natural consequence of scattering.

An electromagnetic wave propagated in a medium may be represented as the sum of the primary wave and the scattered waves. Theoretical calculations show that the superposition of these waves leads to the retardation of the primary wave.

#### Sec. 140. SCATTERING IN A NONHOMOGENEOUS MEDIUM

A substance uniformly distributed, in the sense that we have just discussed, does not scatter electromagnetic waves. Although all portions of this substance create wavelets, secondary radiation is not observed from the sides: no matter what point in the region we chose as our point of observation, it can be rigorously proved that the waves scattered by a uniform substance annul each other due to interference. Since for any particular wavelet there always corresponds another that is exactly opposite in phase, the net effect is complete cancellation.

Now, let us assume that in some limited region the substance has a greater density than in the surrounding medium, i.e., that it has an excess of dipoles per unit volume. Then, all wavelets are annulled except those created by this excess density. As always, the scattered radiation is determined by taking the sum of the wavelet amplitudes, whereby the phase differences between the wavelets arriving at the point of observation must, of course, be taken into account.

Is the situation any different if the density of the scattering region is less, instead of greater, than the surrounding medium? Scattering will cease if matter is added to such a scattering region until the medium becomes homogeneous. Clearly, the net effect is the same if we add a certain amount to a given quantity or subtract this same amount. Hence, the scattering from a region of lower density is equal to the scattering from the missing substance, i.e., the substance required to make the medium homogeneous.

Thus, it is only important that the scattering region have a density distribution of matter that differs from the surrounding medium. Moreover, as regards scattering, the effect of a positive density deviation is indistinguishable from a negative

density deviation of the same magnitude. For example, the scattering from porous glass is the same as the scattering from glass containing randomly distributed inclusions of exactly the same size as the pores.

Due to the large wavelengths of radio waves, scattering will take place in this case only when the nonuniformity of the density occurs on a relatively large scale. For example, in order for the scattering of kilometre waves to be detectable, the extent of the deviations from average density that are intercepted by the waves must be at least several hundred metres. The waves are unable to "detect" smaller inclusions or density gaps.

The scattering of light waves is detectable when disturbances in the distribution of scattering matter are at least of the order of several tenths of a micron. Thus, light waves are unable to detect nonuniformities in the distribution of electrons in a molecule or in the region between two adjacent molecules since these phenomena are restricted to regions whose dimensions are much less than several tenths of a micron. The situation is different as regards X-rays. In this case, the wavelengths are of the same order of magnitude as the dimensions of an atom. Hence, an individual atom appears as an "inclusion in a void".

The scattering of light waves on nonuniformities is a common phenomenon. Nonuniformities in a scattering substance are easily recognised by the external appearance of the medium, i.e., by its turbid appearance. The conditions needed for light scattering prevail in opalescent glass, dust-laden air, etc. In all these cases, there are random disturbances of the density of the substance, whereby the dimensions of the disturbed regions approach the wavelength of light.

As was indicated above, if a particle or nonuniformity scatters as a single dipole, in other words, if its dimensions are no greater than one-tenth to one-twentieth of the wavelength, the scattering intensity as given by the dipole formula (see p. 275) is proportional to the frequency, and inversely proportional to the wavelength, taken to the fourth power. This is the explanation for the following interesting phenomenon. When white light is scattered by a medium with nonuniformities, the medium acquires a blue coloration since the intensity of the scattered blue rays, i.e., those of shortest wavelength, is considerably greater than that of the others. On the other hand, after passing through a scattering medium, white light becomes reddish since the blue portion of the spectrum is impoverished due to greater scattering.

For light waves, not only are turbid media nonhomogeneous. A homogeneous gas or liquid is optically nonhomogeneous due to the presence of density fluctuations. This may be shown by calculations. Light waves scattered in a region whose linear dimension is 0.02 micron, i.e., one-twentieth of the wavelength  $\lambda$ , may be considered to be in phase. There are, on the average, 215 molecules in such a gas volume ( $8 \times 10^{-18}$  cm<sup>3</sup>) under normal conditions. The relative fluctuation on the number of particles according to the laws of statistical physics is  $\frac{1}{\sqrt{N}}$ , i.e., approximately 4 per cent. This is a perfectly perceptible nonuniformity as the scattering of light by air shows.

The blue colour of the sky is due to such scattering. If the atmosphere did not scatter the light of the Sun, the sky would appear black. It should be noted that the colour of the sky is due to the scattering of a relatively small portion of the energy: in a unit volume, about  $10^{-7}$  of the primary wave energy is scattered.

Scattering from density fluctuations in a substance is referred to as molecular scattering since this scattering depends on the molecular structure of the substance rather than its impurities. Investigation of the molecular scattering of liquids is of interest as a method for the determination of certain features of molecular struc-

ture. As regards the nature of scattering, a nonhomogeneous medium in which the regions of deviation from the average density are sufficiently far apart and quite randomly distributed does not differ from a system of randomly distributed scattering centres (Sec. 138). For the most part, however, in discontinuous media (such as fluids and amorphous solid bodies in the case of X-rays, opalescent glass and colloidal systems in the case of light rays, and the atmosphere in the case of radio waves) the interference of waves scattered by neighbouring regions of lower or higher density affects the form of the scattering pattern. This type of interference yields a scattering pattern that significantly differs from the ideal scattering pattern obtained from a single electric dipole.

We have considered scattering of electromagnetic waves by a system of randomly distributed particles, scattering in a perfectly homogeneous medium, and finally, as an intermediate case, scattering in a nonhomogeneous medium. One more important case remains to be discussed, namely, scattering of electromagnetic waves by systems of uniformly distributed centres. Under this case, the following systems will be considered: a diffraction grating for light waves, directional radiators for radio waves and crystals for X-rays.

#### Sec. 141. DIFFRACTION GRATING

A diffraction grating may be constructed using a glass plate coated with a thin layer of aluminium. By means of a special device, uniformly spaced lines are inscribed on this plate with a soft ivory tool. In such a "grating", the nonuniformities (lines) are uniformly distributed, leading to a number of light scattering peculiarities.

We shall be speaking about an optical diffraction grating, but the discussion applies to any regular distribution of nonuniformities and scattering centres and to all electromagnetic waves, i.e., from the shortest to the longest (kilometre waves). The discussion will be restricted to diffraction using parallel rays, which may be realised as indicated in Sec. 137.

If all scattering centres are identical, as is undoubtedly the case in an optical diffraction grating, the diffraction pattern may be determined in the following manner.

Consider the amplitude of the wave emerging at an angle  $\varphi$  to the incident wave. The total amplitude is equal to the sum of the amplitudes of the waves scattered by the individual centres. If the waves from the individual centres arrived at the point of observation in phase, the total amplitude would be equal to the product of the number of centres  $N$  and the amplitude of an individual centre  $f$ . However, the wave from each centre is displaced in phase relative to the wave from an adjacent centre, and it may be assumed that the magnitude of the displacement is the same in each case. Waves from different centres interfere with each other and the resultant intensity is equal to  $Lf^2$  rather than  $Nf^2$ , where  $L$  is a quantity greater than  $N$  for the directions in which the waves reinforce each other and less than  $N$  for the directions in which they arrive, in the main, out of phase and annul each other.

The directions for which the waves from all centres reinforce each other may be easily determined with the aid of Fig. 158. The path difference between waves emanating from corresponding points of adjacent centres is equal to  $a \sin \varphi$ . If this path difference is equal to a whole number of wavelengths, the waves reinforce each other:  $a \sin \varphi = n\lambda$  (maximum condition). As may be seen, there are several such directions. If the wave impinging on the grating is not monochromatic, the grating resolves the wave into its spectrum. Moreover, there will be sev-

eral spectra rather than just a single spectrum. The number  $n$  appearing in the above equation is called, therefore, *the order of the spectrum*.

The number  $n$  may equal zero (in the case of an undeflected ray) or be negative. The first and minus first, second and minus second, etc. spectra will be identical for a simple geometry of the experiment (a plane wave at right angles).

Since it is somewhat cumbersome to calculate the intensity distribution of a scattered wave, let us merely confine ourselves to one important problem, namely, the determination of the width of a diffraction maximum. We are interested in determining how rapidly the intensity of a diffraction maximum, which occurs at an angle  $\varphi$  satisfying the equation  $a \sin \varphi = n\lambda$ , decreases. Does one maximum immediately pass over into the next or is there a large interval between maxima? The case of a grating consisting of a large number of scattering centres (lines) is of considerable practical importance. Imagine the grating divided into two

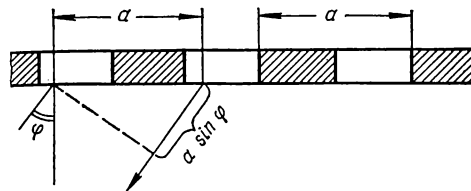


Fig. 158

parts. Now, compare a pair of rays coming from the first centre and the  $(\frac{N}{2} + 1)$  st centre, the second centre and the  $(\frac{N}{2} + 2)$  nd centre, etc. At maximum reinforcement of the waves, the path difference between such pairs of rays is equal to  $\frac{N}{2} n\lambda$ . If the path of the rays is slightly changed, the rays being inclined so that the path difference increases by  $\frac{\lambda}{2}$ , maximum reinforcement of the superimposed waves is replaced by total annulment. The first wave cancels the  $(\frac{N}{2} + 1)$  st, the second cancels the  $(\frac{N}{2} + 2)$  nd, etc. Exact calculations show that for positions further away from the maximum the intensity remains practically equal to zero until the angle of inclination  $\varphi$  approaches the position of the next maximum.

The angle at which maximum diffraction of  $n$ -th order occurs is given by the formula  $\sin \varphi = \frac{n\lambda}{a}$ .

If we designate the half-width angle of the maximum by  $\Delta\varphi$ , then for the angle  $(\varphi + \Delta\varphi)$  we may write the condition

$$\frac{N}{2} a \sin (\varphi + \Delta\varphi) = \frac{N}{2} n\lambda + \frac{\lambda}{2}.$$

Hence,

$$\sin (\varphi + \Delta\varphi) = \frac{n\lambda}{a} + \frac{\lambda}{Na}$$

or

$$\sin (\varphi + \Delta\varphi) - \sin \varphi = \frac{\lambda}{Na}.$$

The distance between two successive maxima is determined by the expression

$$\sin \varphi_2 - \sin \varphi_1 = \frac{\lambda}{a}.$$

We see that the half-width of a line is, roughly speaking,  $\frac{1}{N}$  of the distance between maxima. When  $N$  is large, i.e., in the case of grating consisting of a large number of scattering centres, the diffraction lines are extremely narrow and the

resolution in the spectrum obtained from the grating is very fine. Imagine, for example, that light containing two close waves,  $\lambda$  and  $\lambda + \delta\lambda$ , impinges on a grating. For simplicity, assume we are concerned with scattering at angles less than  $20^\circ$ ; hence,  $\sin \varphi \approx \varphi$ . Then, in the  $n$ -th order these two lines are displaced by the angle  $\delta\varphi$ , which, as can be seen from the condition  $\varphi \approx \sin \varphi = \frac{n\lambda}{a}$ , is approximately equal to  $\frac{n}{a} \delta\lambda$ . The width of the maximum for each wave may be determined from the equation

$$\sin(\varphi + \delta\varphi) - \sin \varphi \approx \delta\varphi = \frac{\lambda}{Na}$$

Evidently, these two lines may be distinguished (in optics, we say **resolved**) if

$$\frac{n}{a} \delta\lambda \geq \frac{\lambda}{Na}.$$

The expression  $\frac{\lambda}{\delta\lambda} = nN$  is an indicator of the *resolving power* of the grating.

*Example.* In a good diffraction grating, the distance  $a$  between lines is  $a \sim 10^{-3}$  mm and the number of lines  $N$  is equal to 100,000. Then, the resolving power for the spectrum of second order is  $\frac{\lambda}{\delta\lambda} = nN = 200,000$ . This means that for  $\lambda = 6,000 \text{ \AA}$ , for example, two lines whose wavelength difference is  $0.03 \text{ \AA}$  may be resolved.

Let us consider the intensity of a diffracted beam. The waves directed towards a maximum point act in phase. If  $f$  is the amplitude of the wave scattered by one centre, the total amplitude in the maximum direction is  $Nf$  and the intensity is  $N^2 f^2$ . Thus, the height of a diffraction maximum is proportional to the number of scattering centres squared and, since the width of a maximum is inversely proportional to  $N$ , its area (i.e., the integral of the maximum intensity) is proportional to  $N$  taken to the first power. If different maxima are compared, it will be seen that the ratio of their heights (or, what amounts to the same, their areas) depends on the value for these directions of the amplitude  $f$  of the scattering from one centre.

Thus, the period of a grating determines the locations of the maxima, while the form (in the broad sense of the word) of a line or scattering centre determines the intensity of the maxima.

Let us assume the angles  $\varphi_1, \varphi_2, \varphi_3$ , etc., are determined by the period of a grating. Scattered rays occur only at these angles. But what will be the intensity of these rays for the first, second, etc., orders of diffraction? This depends on the amplitude values of one scattering centre for these scattering angles. Thus, for example, the amplitude  $f$  may be a maximum at the angle  $\varphi_2$ . Then, the second order diffraction will be represented by an intense line. If at the angle  $\varphi_3$ , the amplitude  $f$  is close to zero, then the third order line will not appear in the diffraction spectrum, etc. This is illustrated in Fig. 159, which shows the diffraction spectra and scattering factors  $f$  of one centre (dotted curves) for two different grating arrangements.

These principles form the basis of any study of structure by means of diffraction spectra. The distance between diffraction lines enables us to determine the period of the grating—assuming, of course, that the wavelength is known—and the intensity of lines of different order enables us to determine the structure of a scattering centre.



*Example.* Consider a diffraction grating for which  $a = 3 \times 10^{-3}$  mm and  $N = 1,000$ . Assume a parallel monochromatic beam of light ( $\lambda = 5,000$  Å) impinges on this grating. Diffraction maxima are visible at the angles given by  $\sin \varphi_n = \frac{n\lambda}{a} = \frac{n}{6}$ , and the width of a diffraction maximum

is equal to  $2 \frac{\lambda}{Na} = \frac{1}{3,000}$ . The obtained results are valid for any scattering centre form. To calculate the relative intensity of the diffraction maxima, it is necessary to consider specific scattering centres. Let us examine two cases:

1. The scattering centres are single strips of width  $b = \frac{1}{4}a = 0.75 \times 10^{-3}$  mm (Fig. 159a). In Sec. 137, we obtained a formula for the intensity of a wave diffracted at a slit. The magni-

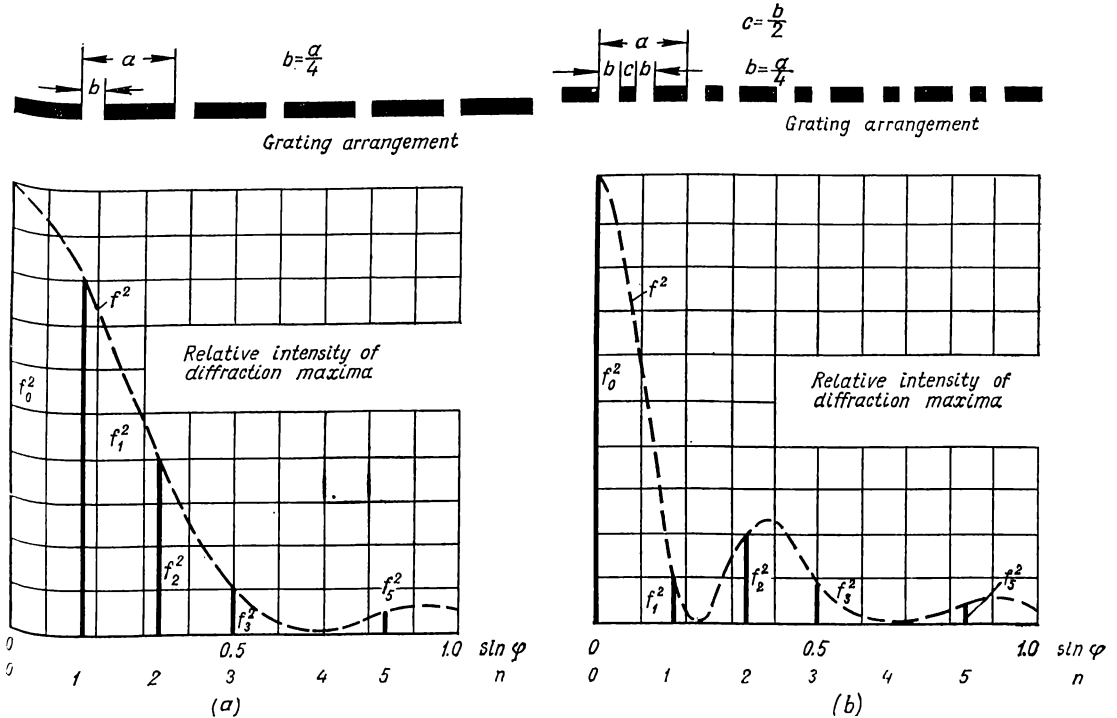


Fig. 159

tude of the intensity is proportional to the amplitude squared of the scattering from one strip:  $f^2 \sim \frac{\sin^2 u}{u^2}$ , where  $u = \frac{\pi b}{\lambda} \sin \varphi$ . The intensity of the  $n$ -th diffraction maximum is determined by the magnitude of  $f_n^2$  in the direction  $\varphi_n$ , which is determined, in turn, from the equation  $\sin \varphi_n = \frac{n\lambda}{a}$ . If  $f_0^2$  is taken as 100, the other intensities have the following values:

$$f_1^2 = 80; \quad f_2^2 = 40; \quad f_3^2 = 8.9; \quad f_4^2 = 0; \quad f_5^2 = 3.2.$$

2. The scattering centres are double strips of width  $b = \frac{1}{4}a = 0.75 \times 10^{-3}$  mm each. The grating period  $a$  is the same as before (Fig. 159b). Clearly, the location and width of the diffraction maxima have not changed. Calculations similar to those performed in Sec. 137 show that for two slits

$$f^2 \sim \frac{\sin^2 u}{u^2} 2 \left\{ 1 + \cos \left[ \frac{2\pi}{\lambda} (b+c) \sin \varphi \right] \right\}.$$

From this expression, the relative intensities of the diffraction maxima are easily determined. Assuming, as before, that  $f_0^2 = 100$ , then

$$f_1^2 = 12; \quad f_2^2 = 20; \quad f_3^2 = 7.5; \quad f_4^2 = 0; \quad f_5^2 = 2.7.$$

#### Sec. 142. DIRECTED RADIATORS OF RADIO WAVES

In certain radio engineering applications, particularly radar, it is required to direct a radio beam into space in such a manner that the transmitted energy is concentrated in a very small solid angle. One solution of this problem is to utilise a linear array of antennas.

In Sec. 141, we saw that for uniform spacing of scattering centres the radiated energy is concentrated in specific directions. If radiators of radio waves are arranged in a single row (see Fig. 160) with a distance  $a$  between adjacent antennas, and if all the antennas are fed synchronously, such a radiator array will in no way differ

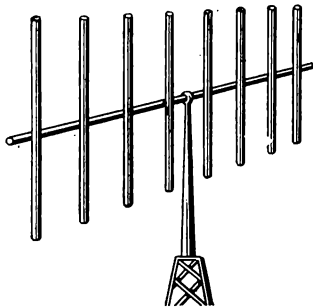


Fig. 160

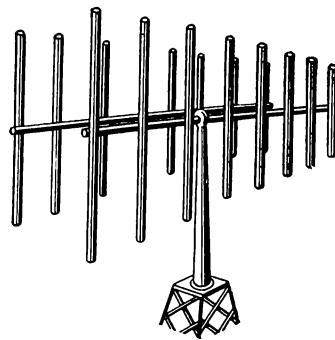


Fig. 161

from a scattering diffraction grating. The fact that we are dealing here with primary waves in no way detracts from the applicability of the discussion in the preceding article to the present case. It is only necessary that it be permissible to consider the radiation from different antennas coherent, which is the case if the antennas are fed synchronously from a generator of oscillations.

If the antenna array is so dense that the distance between adjacent dipoles is less than a wavelength, even first order diffraction is impossible, as the equation  $a \sin \varphi = n\lambda$  shows. Only the zero order remains. This means that there are only two radiation maxima—one forms a  $0^\circ$  angle with the normal to the array and the other a  $180^\circ$  angle. The point made in Sec. 141 about the width of the maximum is valid here too, namely, the larger the total number of radiators, the smaller the solid angle in which the beam intensity reaches an appreciable value.

However, in practice, it is inconvenient to have radiation of equal intensity in two opposite directions. To avoid this, double arrays of dipoles are used (see Fig. 161). The antennas of each dipole pair are separated by a distance of  $\lambda/4$  and a phase difference of  $90^\circ$  exists between the two currents. As a result, one of the two maxima is annulled and all the energy is concentrated into one diffraction maximum. The phase relationships existing for each dipole pair in such an array may be explained as follows. For the "forward" waves: if the waves were propagated synchronously, the path difference between them would be  $\lambda/4$ ; but the antennas do not operate synchronously, i.e., the wave radiated by the "front" dipole lags by  $90^\circ$ , which compensates for its lead due to the path difference. The situation is different for waves propagated "backwards". The displacements due to

the path difference and the phase difference between the currents in the antennas are in the same direction. Hence, the total phase difference is  $180^\circ$ . As a result, there is no radiation in the backward direction.

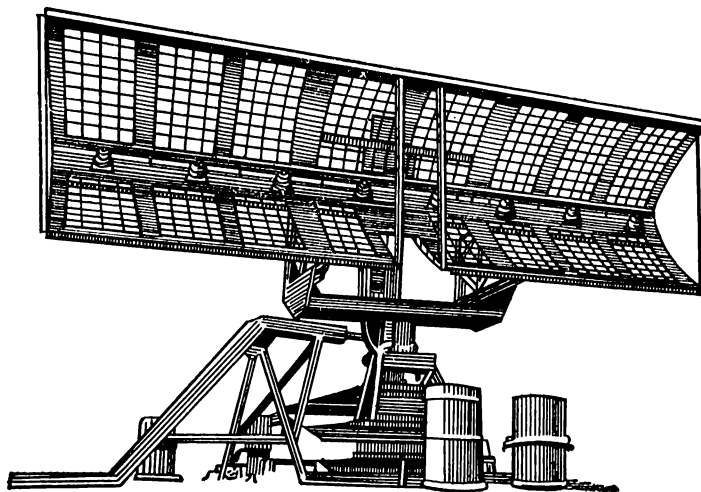


Fig. 162

If the antennas are arranged in a row, a narrow beam may be obtained in the plane perpendicular to the antennas. But if a small beam in space is desired, a more complex system of oscillator elements must be used. That is why radar antennas have such strange appearances (see Fig. 162).

#### Sec. 143. HOLOGRAPHY

A wave scattered by an object (or scatterer) carries rich information concerning the properties of this object. Proceeding from Huygens' principle, we can rigorously prove that the distribution of amplitudes and phases of the wave at its front, at any instant of the process of its propagation characterizes the scattering properties of the object in an exhaustive manner. When the object is being photographed, a part of the information carried by the wave is lost. The blackening of a photographic plate is proportional to the square of the wave amplitude (i.e. intensity) which reached a given point of the plate and does not depend on the wave phase. The photographic plate yields a two-dimensional representation image of a three-dimensional picture.

Thus, it is clear that in photographing a considerable part of information concerning the object is lost.

In 1947 D. Gabor came out with a fundamentally new photographic technique which he called wavefront reconstruction. His method consists in the following.

The images obtained with conventional optical instruments (a photographic camera, a magic lantern, a motion-picture projector, and the like) register only the intensity of a wave, that is, the square of its amplitude, while the wave phase is lost. With Gabor's technique, both the frequency and phase attributes of a wave are recorded as a blurred diffraction pattern; this is the first step. In the second step, the blurred photograph is inserted in an optical device which is capable

of restoring it into a reconstructed image of the primary object. Since the blurred diffraction pattern contains almost all the available information concerning the sample, it is called a *hologram* (from the Greek "holos" for "total" and "gramma" for "record"), and the technique has

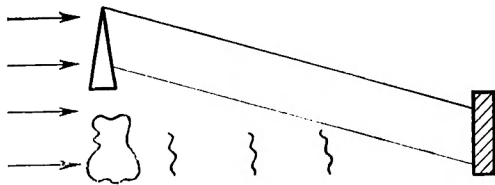


Fig. 162a

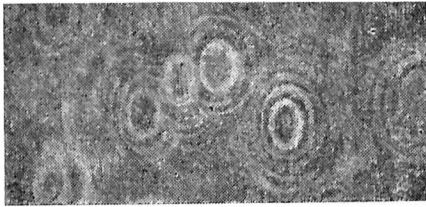


Fig. 162b

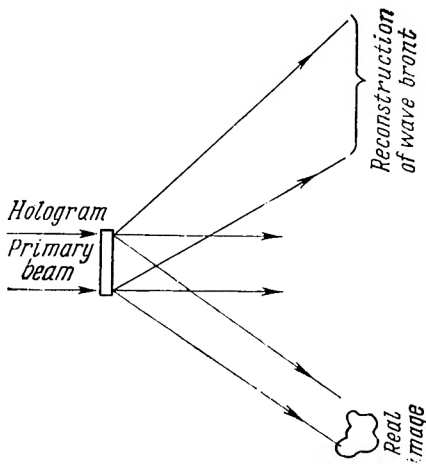


Fig. 162c

finally come to be known as holography.

The holography technique became practically realizable only after the invention of the laser.

Thus, holography is a two-stage process. First, a hologram is taken from an object, and then the hologram is illuminated to obtain a three-dimensional image of the object, in principle, indistinguishable from it.

In other words, a hologram is an interference picture obtained on a photographic plate as a result of interference of the primary and scattered rays.

These rays can be made to interfere in a variety of ways. In the simplest pattern we can cover one part of the plane front of a laser wave with a prism, and the other with a scattering object (Fig. 162, a). The coherent radiation of a laser will enable us to obtain on the photographic plate a certain interference picture whose form will be like one shown in Fig. 162, b. If the scattering object is removed, then the interference picture would have the form of lines parallel to the edge of the prism with a distance between the lines readily expressed in terms of the angle of the prism (see Sec. 134). We may say that this simple interference picture was "modulated" by the distance from the object.

The pattern of alternating black and light strips resembles a diffraction grating. Therefore, it is not difficult to agree that when the hologram is exposed to light, there will appear an undeflected ray, as also the spectra plus and minus of the first order. We can show (but we are not going to dwell on it here (that there will

be no spectra of high orders, and that one of the spectra will represent a diverging reconstructed wave front of the light wave scattered by the object. The other wave will be converging. It will create a real image of the object which can be observed with the same possibilities as the object itself (Fig. 162, c).

At the present time various applications of holography are being discussed. What are the advantages of holography (including colour holography) which is developing so rapidly now? Let us consider some of them.

(1) In an ordinary photograph, every area of emulsion represents a part of the original object. Therefore the information contained in one area is in no way related to that in any other area. The destruction of a part of the photograph leads to the loss of the respective information. In a hologram, each part contains information about the entire object, permitting its reconstruction from any small portion of the hologram, although the reconstructed image may be not so well defined and clear. The situation is similar to using a fragment of a lens for making an image.

Thus, as a medium for data storage, a hologram is more reliable than an ordinary photograph.

(2) As compared with an ordinary photograph, a hologram can store a considerably greater amount of information. While a  $6 \times 9$  mm photograph can hold one printed page, a single hologram plate of the same size can store 100 to 300 such pages, depending on the emulsion quality. Now that the problem of storage media for the rapidly growing printed information is becoming more and more urgent, holography can offer a way out.

(3) Holography can be adapted to fill the need for 3-D colour motion pictures and television.

(4) If the wavelength used for restoring the hologram into a reconstructed image of the primary test object ( $\lambda'$ ) is greater than that used for obtaining the primary diffraction pattern ( $\lambda$ ), the reconstructed image will be magnified, the degree of magnification being proportional to the ratio between the two wavelengths i.e.  $\lambda'/\lambda$ . A remarkably high magnification can be obtained along with an increase in resolving power. However, there is a limit, too. The wavelength of the restoring beam ( $\lambda'$ ) should be a fraction of the spacing between the interference fringes. Otherwise, the emulsion will be an optically homogeneous medium for the restoring wave, and no holographic effect will be obtained.

(5) There is a good deal of attraction in acoustical holography. Coherent acoustic waves are easy to produce, while sound (or ultrasound) is readily propagated in liquids or solids. With them, making a three-dimensional acoustical hologram of an opaque object will present no problem. By restoring the hologram in visible light, we shall be able to see the internal structure of, say, a metal bar, a concrete girder, or a living organism. This is strikingly new opportunity for both engineering and medicine.

A major snag of acoustical holography is the difficulty of catching an acoustical hologram. Yet, research is under way, and some approaches are being tried out.

(6) Holograms can be readily transmitted over any distances, and the image can be reconstructed far away from the object.

(7) Using different wavelengths and inclinations of the primary beam, we can "write" the images of different objects on one and the same plate.

(8) Holographic microscopy is free of the enormous drawback of ordinary microscopy, that is of the necessity to carry out focussing. On receiving a volume hologram, we can examine it, without haste, through a microscope, study in detail all cuts of the object, although the hologram under study refers to a certain fixed instant. Such method of investigation is of special importance for carrying out experiments with living objects.

We have only taken up some of the applications for holography. Many of its potential uses have not yet been discovered theoretically; still fewer have found embodiment in practice. Yet, it will be no wild guess to say that future holds much in store for holography.

# Diffraction of X-Rays by Crystals\*

## Sec. 144. CRYSTALS AS DIFFRACTION GRATINGS

As has been already indicated, a diffraction grating usually consists of a piece of glass on which equally spaced lines are scratched. What is essential here to obtain one of the typical diffraction patterns discussed above? Is it the presence of glass, the shape of the lines, the thickness of the glass, or the width of the "slits"? A careful study of Sec. 141 shows that the essential element is the periodic repetition of the nonuniformities of the scattering substance. Thus, irrespective of the cause of the scattering and the nature of the nonuniformities, as long as these nonuniformities are repeated with a periodicity  $a$ , scattering maxima will occur

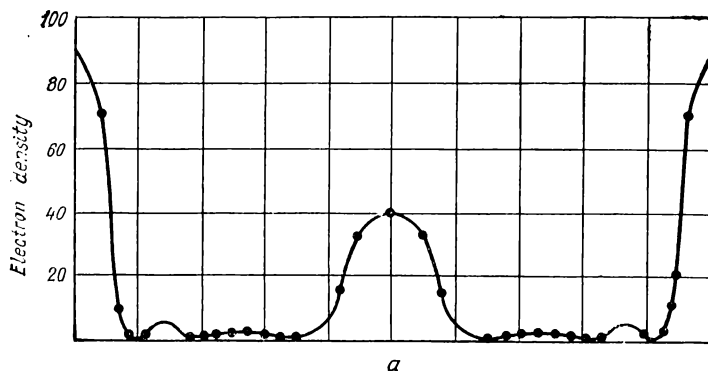


Fig. 163

at angles  $\varphi$  satisfying the equation  $a \sin \varphi = n\lambda$ . Such a pattern is given by lines of any shape scratched on any glass, or slits of any shape made in any screen. It is only necessary that the distribution of matter repeat with a periodicity  $a$ .

To be sure, certain differences in patterns may exist. The intensities of rays diffracted in different orders may differ depending on the shape of the slit. The distribution of substance within a repeating nonuniformity affects the scattering intensity  $f^2$ , which for different orders may have different values.

Now that we have reviewed the results of Sec. 141, let us turn to crystals. The basic feature of crystals distinguishing them from other bodies is the periodic distribution of matter. Along any direction of a crystal, the time-averaged electron density is periodically repeated. In the simplest case, the electron density distribution will appear as shown in Fig. 163. This figure shows the electron density (the number of electrons per cubic Angstrom) along a line parallel to the edge of a cube of rock salt. An electron density maximum corresponds to the centre of an atom. The large maximum corresponds to the centre of a chlorine atom and the small maximum to that of a sodium atom. The pattern repeats itself after every other atom, and the period of the electron density distribution along the line is

\* Before reading Chapters XXII and XXIII, it is recommended that Chapter XXXII be perused.

equal to  $5.6 \text{ \AA}$ . This is the pattern of the distribution obtained along a specific line. Along a slightly displaced parallel line, the distribution will be different.

However, a crystal is a three-dimensional formation, and the repeating element is a three-dimensional cell. The electron density distribution of a cell cannot be illustrated graphically, but it is sufficient to know that the cells repeat in space. The similarities and differences between a crystal and a diffraction grating are evident. A crystal is a three-dimensional diffraction grating in which the nonhomogeneous element repeats regularly in three-dimensions rather than along a line. The role of "slit", i.e., repeating nonuniformity, is played by the unit cell of the crystal.

Let us determine the nature of the diffraction pattern created by a crystal.

X-rays are scattered by electrons. The nonuniformities in the electron density are of such a nature that wavelengths of the order of  $1\text{--}2 \text{ \AA}$  yield perceptible diffraction. In order to determine the direction of the diffracted rays, the wavelets coming from all the cells must be added. Of course, the amplitudes of these wavelets for a given direction are the same. The difficulty arises in taking account of the phase differences between wavelets scattered by individual cells. These wavelets must be added for every direction and the directions in which the wavelets reinforce each other to a maximum extent determined.

The problem can be solved in various ways since various summation sequences are possible. For example, first wavelets from the cells along edge  $a$  may be added, then wavelets from all columns in the plane  $ab$ , etc. But we shall use a much simpler method. This is the method proposed by the founders of X-ray structural analysis, the two Braggs—father and son, British scientists. (The same idea was proposed independently by the Russian crystallographer Vulf.) In a crystal, parallel planes can always be passed through the lattice nodes in numerous ways; the crystal is composed of the layers between such successive planes. Let us construct a normal to these layers and imagine the electron density projected onto the direction of the normal. Clearly, a periodic electron density distribution exists along the normal. The period  $d$  is appropriately called the interplanar distance.

The condition for maximum reinforcement of waves scattered by the cells of one layer is that the angle of incidence equals the angle of reflection. This conclusion is based on Huygen's principle, for only when the above condition is satisfied are the scattered waves propagated in phase, and thus reinforced. Waves of successive layers reinforce each other when certain additional conditions are met. Fig. 164 shows that the path difference between rays "reflected" from the corresponding elements of two adjacent layers is equal to  $2d \sin \theta$ . Thus, diffracted rays are obtained when the condition

$$2d \sin \theta = n\lambda$$

is satisfied.

A diffracted beam is obtained when a system of planes may be found, among the countless such systems into which the crystal may be divided, which satisfies the condition  $2d \sin \theta = n\lambda$ . Of course, this requirement may simultaneously be

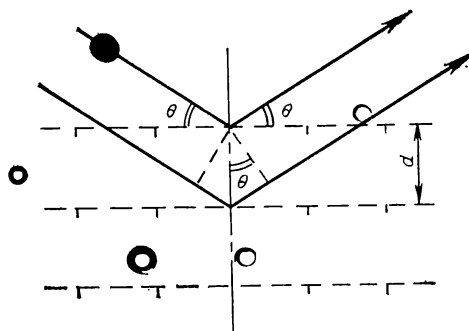


Fig. 164

met by several systems of planes. However, the more likely situation is that diffraction does not occur for an arbitrary direction of a monochromatic beam. Hence, in order to observe diffraction, the crystal must be turned until a suitable angle  $\theta$  is found.

*Example.* The interplanar distance in a calcite crystal is equal to  $3.029 \text{ \AA}$ . In X-ray structural analysis, radiation from a copper anode is often used. Since this radiation has a wavelength of  $1.54 \text{ \AA}$ , the diffraction maximum of first order occurs at  $\theta = \arcsin \frac{\lambda}{2d} \approx 14^\circ 40'$ .

#### Sec. 145. DETERMINATION OF THE PARAMETERS OF A CRYSTAL CELL

Having experimentally determined the angles  $2\theta$  formed by rays diffracted from a crystal, one can determine, if  $\lambda$  is known, the interplanar distances, and hence the structural period in any direction.

If the lattice is cubic, it is described by one parameter, i.e., the edge of a cube. A rhombic lattice is described by three mutually perpendicular periods  $a$ ,  $b$  and  $c$ .

Rays may be reflected from any system of planes, including the planes parallel to the main lattice faces  $ab$ ,  $bc$  and  $ac$ , by appropriate placement of the crystal relative to the beam. A series of such measurements suffices in all cases to fully

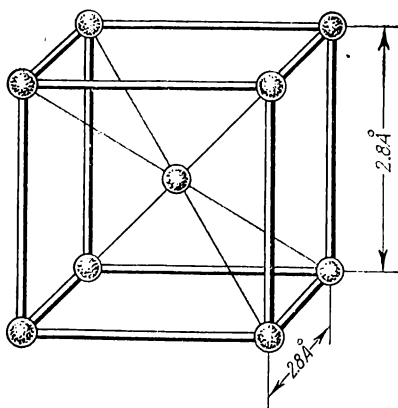


Fig. 165

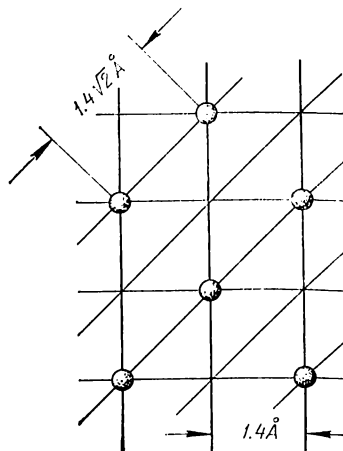


Fig. 166

“probe” the lattice structure, i.e., measure the length of the edges of a unit cell and, for the case of low-order symmetry, determine the angles between them.

These measurements are quite interesting. Having determined the volume  $V$  of a crystal cell and the density  $\delta$  of the substance, the mass of substance in a cell may be directly calculated from  $M = V\delta$ . The molecular weight of a cell is then obtained by dividing this quantity by the mass of a hydrogen atom:  $m_{H} = 1.66 \times 10^{-24} \text{ g}$ . Of course, the number of molecules in a cell must be an integer. Moreover, in many cases, symmetry considerations restrict the possible number of molecules. For example, in a rhombic cell, the number of molecules cannot be less than four. Thus, cell measurements are of great importance in the determination of the molecular weight of a substance.

In the simplest cases, the determination of a unit cell suffices to determine the structure. A unit cell of iron is illustrated in Fig. 165. For such a cell, the distances between atoms are as indicated. These data are obtained by simply measuring the



interplanar distances. The investigator reasons as follows: a crystal of iron is cubic. Let us measure the fundamental interplanar distance for the system of planes parallel to the side of a cube. This yields the value  $1.4 \text{ \AA}$ . Now, let us calculate the number of iron atoms in a cube whose edge is  $1.4 \text{ \AA}$  long. The mass of an iron atom is  $m_{\text{Fe}} = 92.6 \times 10^{-24} \text{ g}$  and the density of iron is  $\delta_{\text{Fe}} = 7.88 \text{ g/cm}^3$ . Thus, a cube of volume  $(1.4)^3 \times 10^{-23} \text{ cm}^3$  contains a mass of  $(1.4)^3 \times 10^{-24} \times 7.88 = 21.7 \times 10^{-24} \text{ g}$ . But this mass is one-fourth that of an iron atom. Hence, the edge of a unit cell of iron is greater than  $1.4 \text{ \AA}$ . Let us assume that it is twice this length. Then, the edge of a cube is  $2.8 \text{ \AA}$  and we obtain 2 atoms per cell.

Since the crystal is cubic and possesses an axis of symmetry of fourth order, the second atom can be located only at the centre of a unit cube. Let us check the validity of the assumption of two atoms per cell. If the assumption is correct, the diagonal interplanar distance shown in Fig. 166 should equal  $1.4\sqrt{2} \text{ \AA}$ . This is the value obtained experimentally. Hence, the above model of the structure is correct.

For many metals, alloys and simple salts whose formulas are of the type  $AB$ , such simple reasoning is quite often sufficient to determine the interatomic spacing. If there are many atoms in a cell and the shape of the cell is not cubic, the structure may be determined only by utilising beam intensity data in addition to data on the geometry of the diffraction pattern.

#### Sec. 146. INTENSITY OF DIFFRACTED BEAMS

Analogously to the case of a linear grating, the intensity of a beam diffracted by a crystal is equal to  $N^2 F^2$ , i.e., it is proportional to  $F^2$ , the square of the amplitude of the wave scattered by a unit crystal cell in a given direction, and to  $N^2$ , the square of the number of unit cells in the illuminated volume.

The quantity  $F^2$  is uniquely related to the crystal structure, i.e., to the nature of the electron density distribution in the cell.

As has been already indicated, the quantity  $F^2$  for a given diffracted beam and system of "reflecting" planes depends on the projection of the electron density on the normal. By means of parallel planes, let us divide a crystal layer into infinitely thin layers  $dz$ . If  $\rho$  is the electron density of a cell, then  $\rho dz$  is the number of electrons in the layer  $dz$ . All the electrons of a thin layer are scattered in phase and yield the wave  $\rho dz \cos(\omega t + \varphi)$ . Hence, the wave amplitude of a cell is expressed by

$$F = \int_0^d \rho \cos(\omega t + \varphi) dz.$$

The integral is taken over a single period (the interplanar distance); the values of  $\rho$  and  $\varphi$  for each  $z$  differ.

We shall not concern ourselves with the problem of determining the electron density of a crystal at all points of a unit cell. In all cases the basis of the solution is the above equation. It enables us to calculate the amplitude of any diffracted beam if the electron density of the crystals is known. Of interest, however, is the converse problem—the determination of the electron distribution in a crystal from the experimentally determined intensities of diffracted beams. This problem has been solved for very complex crystals containing hundreds of atoms per unit cell.

In the case of anthracene crystals, the intensity of about 600 diffracted beams has been measured. Using these data, the values of the electron density at all points

of such a cell were determined. Fig. 167 illustrates the electron density for a cross-section through the centres of atoms of an anthracene molecule. (The method adopted to indicate the electron density is that used by topographers to indicate elevation. Electron density contour lines on the electron density diagram correspond

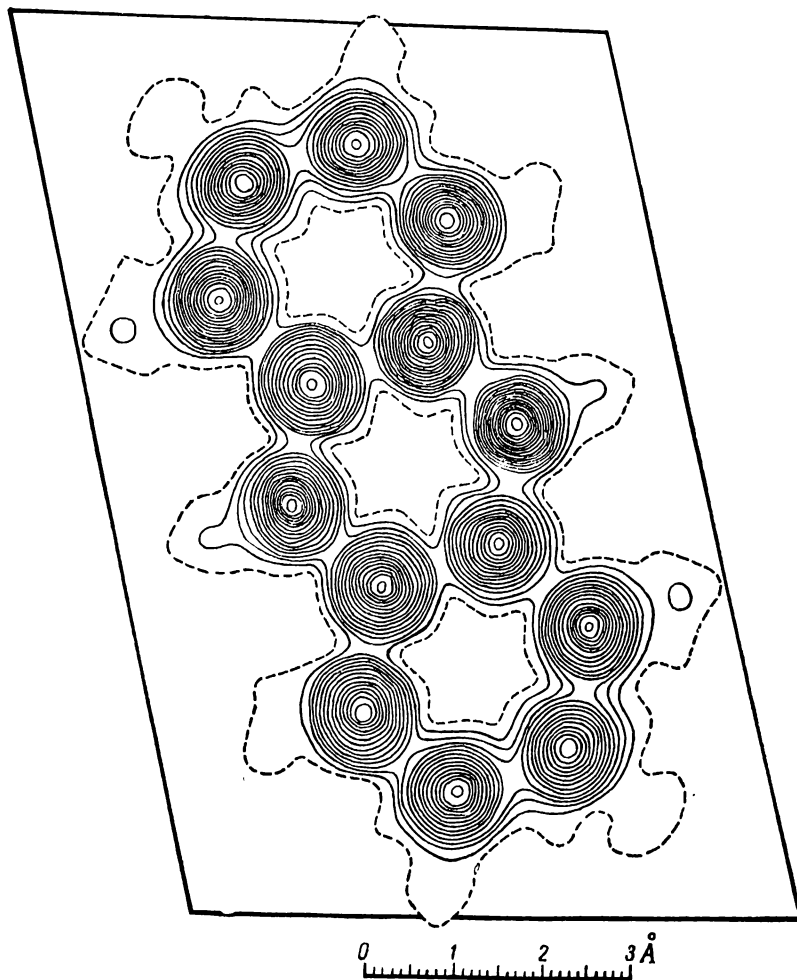


Fig. 167

to elevation contour lines on a topographical map.) The fourteen distinct peaks represent fourteen hydrogen atoms. Experiments show that the distance between adjacent peaks is  $1.4 \text{ \AA}$ . The height of an electron density maximum is proportional to the number of electrons in the atom. Carbon atoms have six electrons each and hydrogen atoms one. Not all of the ten peaks corresponding to the centres of the ten hydrogen atoms contained in an anthracene molecule are fully sketched in the diagram. The chemical formula of anthracene is  $\text{C}_{14}\text{H}_{10}$ .

#### Sec. 147. METHODS OF X-RAY ANALYSIS

The measurement of the angles  $2\theta$  formed between the diffracted beams and the beam incident on a crystal, as well as the intensities of the diffracted beams, may be accomplished using an ionisation chamber (see p. 524) or a photographic method.

A photographic film on which the traces of many diffracted beams are simultaneously recorded is called a roentgenogram.

But how can the traces of several beams be obtained on a single film when, as has been already indicated, the condition  $n\lambda = 2d \sin \theta$  in all likelihood will not be fulfilled even once for an arbitrary orientation of the crystal relative to the beam? This can be accomplished in three ways:

(1) by rotating the crystal, thereby setting various systems of planes in a reflecting position;

(2) by illuminating the crystal with a continuous spectrum of wavelengths in a band sufficiently broad so that almost every system of planes finds a "suitable" wavelength in the spectrum; and

(3) by obtaining a roentgenogram of a powder, since in a powder the diffraction condition for any  $d$  is always fulfilled for some crystals.

The first method is known as the crystal rotation method, the second as the Laue method, after the German physicist whose name is associated with the discovery of beam diffraction, and the third as the powder of Debye method, after the scientist who proposed this method. The Laue method has extremely limited application. In practice, the rotation method is used for the investigation of crystal structure, i.e., for the determination of arrangement of atoms and the Debye method is used for special problems arising in connection with the investigation of fine crystalline substances.

The purpose of a rotating roentgenogram is to gather data on one film regarding existing interplanar distances and intensities of the respective beams. However, it is also necessary to know the orientation of the system of planes relative to the crystal axes. This requires knowledge not only of the location of a particular spot on the film, but also of the instantaneous orientation of the crystal when it arose. In order for the roentgenogram to provide this information as well, the film is displaced during filming. Such filming methods are known as roentgengoniometric methods.

Fine crystalline substances are used much more often than monocrystals for the investigation of crystal structure by means of X-rays. Fig. 168 illustrates how a roentgenogram is obtained by the Debye or powder method.

Let us direct our attention to a specific system of planes separated from one another by a distance  $d$  and having a normal  $n$ . Assume that the normals  $n$  of the crystals are directed in all directions. A beam "reflected" from a plane is not produced by reflection from all the crystals, but only from those whose planes are at an angle  $\theta$  to the beam, where  $\theta$  satisfies the condition  $2d \sin \theta = n\lambda$ . Accordingly, the normals of these crystals are at an angle of  $90^\circ - \theta$  to the primary beam. All crystals whose normals lie on the cone having an apex angle  $2(90^\circ - \theta)$  are in a reflecting position for planes separated from one another by the distance  $d$ . Therefore, the "reflected" beams also form a cone, and this cone has an apex angle of  $4\theta$ . Ring patterns are obtained when these cones are intersected by a photographic film or plate.<sup>2</sup>

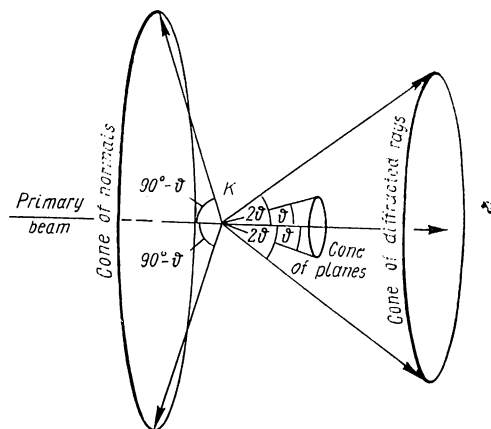


Fig. 168

If we are concerned with values of  $\vartheta$  not exceeding  $20\text{--}25^\circ$ , the roentgenogram may be taken on a flat plate. In this case we obtain a system of concentric rings. However, if information on all interplanar distances is desired, and scattering over the entire interval of angles possible is to be analysed, cameras with cylindrical film are used. In this case, the height of the film is reduced, so that only parts of the rings are photographed.

If for some problem or other it is important to ascertain the interplanar distances more accurately, we usually resort to "rear" filming, i.e., an arrangement whereby diffraction cones having apex angles close to  $360^\circ$  are filmed.

In determining the scattering angle  $\vartheta$  from the film, a measurement error  $\Delta\vartheta$  inevitably arises. Let us see how the magnitude of this error is reflected in the determination of the magnitude of the interplanar distance. After differentiating the diffraction condition  $2d \sin \vartheta = n\lambda$ , we obtain  $\left| \frac{\Delta d}{d} \right| = \cot \vartheta \Delta \vartheta$ . It is seen that the accuracy in measuring the interplanar distance rapidly increases as the angle  $\vartheta$  approaches  $90^\circ$ . The diffraction angle may be easily measured with an accuracy of the order of  $0.1^\circ$ . Hence, the above equation shows that along lines for which the angle  $\vartheta$  is equal to  $65^\circ$ ,  $75^\circ$  and  $85^\circ$  the interplanar distance may be measured with an accuracy of 0.13%, 0.08% and 0.05%, respectively. Using special cameras, the method of rear filming yields very good results, enabling interplanar distances to be determined with an accuracy of  $0.00001 \text{ \AA}$ . For best results, the radiation wavelength is selected so that the scattering angle approaches  $90^\circ$ .

All three types of Debye filming are widely used in the investigation of the structure of matter. Every substance yields a specific system of lines that is characteristic only of it. A phase transformation results in the replacement of one system of lines by another and the new phase may also be determined from the roentgenogram. Phase analysis is one of the most important applications of these films.

Now, let us assume that the crystals of the substance have a favoured orientation. In this case, the normals  $n$  of one or another system of planes no longer assume all directions in space and are not uniformly distributed on the cone of normals shown in Fig. 168. But if the cone of normals is not complete, the same holds true for the cone of diffracted rays. Hence, instead of solid rings, broken rings appear on the film. Such rings indicate the presence of a favoured crystal orientation or structure. Using the Debye method, a detailed study may be made of the nature of the preferred crystal orientation arising, as a rule, for various kinds of plastic deformations.

The crystal dimensions also affect the form of the picture. If the crystals are very large, the roentgenogram ring is not solid, i.e., it breaks down into dots, each of which is due to "reflection" from an individual crystal. If the crystals are very small (of the order of  $10^{-5} \sim 10^{-6} \text{ cm}$ ), then, in accordance with the theory, the lines begin to become smeared. Based on this knowledge, methods of estimating the average dimension of crystals have been developed.

For purposes of analysis, another problem arises in a number of cases, namely, to construct the apparatus in such a manner that the X-ray spectrum radiated by a substance may be investigated. As will be seen below (p. 393), every substance can produce a characteristic X-ray spectrum. The fact that the spectra of different kinds of atoms differ significantly from one another may be utilised for the conduction of qualitative and quantitative analyses. For this purpose, the face of a large crystal, for which the interplanar distance is known, is placed in a "reflecting" position. By rotating this crystal and measuring the diffraction intensity for each angle, the wavelengths present in the spectrum of the substance under investigation, and their intensities, may be determined from the values of the angle  $\vartheta$ .

## Double Refraction

### Sec. 148. ANISOTROPIC POLARISABILITY

When a substance is placed in an electric field, the bound charges, i.e., the electron clouds of the molecules, are displaced from their equilibrium positions and form electric dipoles. In discussing the polarisation of a dielectric (p. 191), we spoke of the displacement of electric charges along lines of force. But clearly the situation may be entirely different in actual molecules. If an atom has a spherically symmetric electron cloud, the latter is indeed displaced to the same extent in any direction. The polarisability of such an atom is said to be isotropic. It is evident that even a diatomic molecule cannot be isotropic, for the nature of the electron bond parallel and perpendicular to the line connecting these two atoms, and therefore the polarisability in each direction, is not the same. The magnitude of the polarisability, and hence of the displacement (dipole moment), will vary depending on the direction of the electric vector relative to the axis of the molecule.

It is also significant that the displacement direction and the electric force, direction do not, generally speaking, coincide. This may be illustrated by a mechanical model—a bead attached to two perpendicular springs of different elasticity. Assume, for example, that a force acts at a  $45^\circ$  angle to the bonds. Different forces will then act along the bonds. Let us assume that the elasticity of the horizontal spring is three times greater than the elasticity of the vertical spring. The vertical displacement will then be three times greater than that of the horizontal, and the displacement vector will form a large angle with the electric force vector. In exactly the same manner, the induced dipole moment forms an angle with the direction of the field intensity  $E$ .

Almost all molecules possess anisotropic polarisability, but this polarisability is by no means manifested in all cases. It does not manifest itself in liquids, amorphous bodies and gases. Thus, for each molecule whose dipole moment is inclined to the "left", there is a conjugate whose dipole moment is inclined to the "right". The resultant dipole moment of such a pair of molecules, and hence the dipole moment per unit volume, i.e., the polarisation vector  $P$ , is directed along the vector  $E$ .

However, even in the case of crystals, where the arrangement of molecules is uniform, the anisotropic polarisability of the molecules is not necessarily manifested. Cubic crystals possess isotropic polarisability. Here, the equation  $P = \alpha E$  remains valid, just as for isotropic bodies. To make this clear, let us consider molecules whose electrons may be displaced only in a single direction. The symmetry of a cubic crystal is such that molecules whose axes of polarisability form right angles may always be found in the crystal. Let us consider three such molecules whose axes of polarisability lie along the  $x$ ,  $y$ ,  $z$  axes of a Cartesian system of coordinates. When an arbitrarily directed field  $E$  is applied, these molecules become polarised and produce dipoles of moments  $\beta E \cos \varphi_1$ ,  $\beta E \cos \varphi_2$  and  $\beta E \cos \varphi_3$ , since the moment is proportional to the projection of  $E$  on the polarisation direction. The resultant of these three dipole moments is obtained by vector addition. But  $\varphi_1$ ,  $\varphi_2$ ,  $\varphi_3$  are the angles formed by the vector  $E$  with the coordinate axes. Hence, the magnitude of the resultant moment is  $E \sqrt{\beta^2 (\cos^2 \varphi_1 + \cos^2 \varphi_2 + \cos^2 \varphi_3)}$   $= \beta E$  and the direction is parallel to  $E$ . Taking the summation for all the mole-

cules, we arrive at the same conclusion as regards the polarisation vector  $\mathbf{P}$  and the polarisability  $\alpha$  per unit volume.

Now, let us consider crystals having one main axis, i.e., crystals with tetragonal and hexagonal syngony. To be specific, let us consider the former, i.e., crystals in which each molecule has three identical molecules related to it via an axis of symmetry of the fourth order. Assume, moreover—again for purposes of simplicity—that the molecule can be polarised only along one axis. Let us direct our attention to the four molecules whose axes of polarisability form an angle  $\varepsilon$  with the main axis (see Fig. 169). How do these molecules behave in electric fields of various directions? If the vector  $\mathbf{E}$  is directed along the main axis, the polarisation is proportional to  $\cos \varepsilon$ . Moreover, in view of the symmetry of the arrangement, the resultant dipole moment of these molecules is parallel to  $\mathbf{E}$ ; hence, the polarisation vector is also parallel to  $\mathbf{E}$ :

$$\mathbf{P} = \alpha_{\parallel} \mathbf{E},$$

where  $\alpha_{\parallel}$  is the polarisability created by all the molecules for this direction of the field.

In the projection onto the plane perpendicular to the main axis, the polarisability axes form right angles with each other. Therefore, the result obtained for a cubic crystal is valid here, namely, the polarisability is the same for all directions of  $\mathbf{E}$  in the plane perpendicular to the main axis. If  $\mathbf{E}$  is perpendicular to the main axis, the polarisation vector  $\mathbf{P}$  is again parallel to  $\mathbf{E}$ :

$$\mathbf{P} = \alpha_{\perp} \mathbf{E},$$

While the polarisability of the molecules along the axis is proportional to  $\cos \varepsilon$ , the polarisability of the molecules in the direction perpendicular to the axis is proportional to  $\sin \varepsilon$ . This means that  $\alpha_{\parallel}$  and  $\alpha_{\perp}$  are different.

What is the situation when the field  $\mathbf{E}$  is inclined to the main axis of the crystal? In view of the difference in the polarisabilities  $\alpha_{\parallel}$  and  $\alpha_{\perp}$ , the vector  $\mathbf{P}$  can no longer coincide with the direction of the field, and the value of  $\alpha$  will also be different. Knowing  $\alpha_{\parallel}$  and  $\alpha_{\perp}$ ,  $\alpha$

may be calculated for any direction. We shall not explain how this is done in the general case, but merely cite a numerical example.

For a crystal of Iceland spar (calcite),  $\alpha_{\perp} = 0.139$  and  $\alpha_{\parallel} = 0.095$ . Assume that the vector  $\mathbf{E}$  forms a  $30^\circ$  angle with the plane perpendicular to the main axis and that it is directed as shown in Fig. 170. The polarisation vector in the indicated plane is then

$$P_{\perp} = \alpha_{\perp} E_{\perp} = 0.139 \times E \cos 30^\circ = 0.120E.$$

The polarisation vector perpendicular to this and parallel to the main axis is

$$P_{\parallel} = \alpha_{\parallel} E_{\parallel} = 0.095 \times E \sin 30^\circ = 0.0475E.$$

Therefore, the angle between the resultant polarisation vector  $\mathbf{P}$  and the plane is  $\arctan \frac{0.0475}{0.120} \approx 21^\circ 40'$ . This means that the angle between  $\mathbf{P}$  and  $\mathbf{E}$  is  $\sim 8^\circ 20'$ .

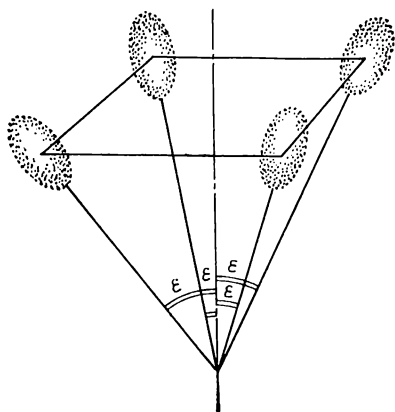


Fig. 169

The magnitude of the polarisation vector is  $P = \sqrt{P_{\perp}^2 + P_{\parallel}^2} = 0.129E$ , i.e., in this case the polarisability  $\alpha$  is equal to 0.129.

If  $E$  is directed at a  $45^\circ$  angle to the main axis, the polarisability  $\alpha$  decreases even more relative to  $\alpha_{\perp}$  and becomes equal to 0.120. The angle between  $E$  and  $P$  is then

$$45^\circ - \arctan \frac{0.095}{0.139} \approx 10^\circ 30'.$$

The direction of the vector  $E$  in Fig. 170 is such that the angles formed by this vector with the polarisability axes of the molecules differ. Thus, the angles are greater for the pair of molecules on the left. As a result, the right pair of molecules is polarised to a greater extent.

The fact that the polarisability of a crystal possesses an axis of symmetry of the fourth order does not signify symmetry in the contributions of individual

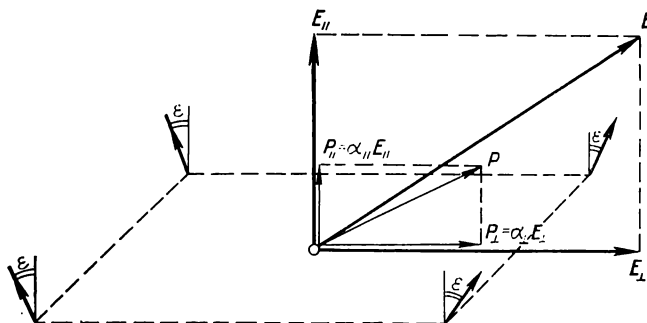


Fig. 170

molecules to the magnitude of the polarisability for an arbitrary field direction.

Thus, the values of the polarisabilities differ for different directions. This has important consequences: the polarisability is uniquely related to the dielectric constant; but  $\epsilon$  determines the index of refraction (see Sec. 125;  $\epsilon = n^2$ ) and hence the wave propagation velocity in the crystal; as a result, the electromagnetic wave is propagated in the crystal with different velocities in different directions.

Tetragonal and hexagonal crystals (in optics they go under the heading “uniaxial”) possess the following characteristic: all orientations obtained by successive rotation about the main axis are optically equivalent. Crystals having a lower order of symmetry do not possess this feature.

Uniaxial crystals have a main direction and, perpendicular to it, a main plane. If the vector  $E$  points in this direction or lies in this plane, then  $P \parallel E$  (and, therefore,  $D \parallel E$ ). Analysis shows that in other crystals only three main mutually perpendicular directions in which  $P \parallel E$  may be distinguished.

#### Sec. 149. PROPAGATION OF LIGHT IN UNIAXIAL CRYSTALS

**Division of a Light Field into Two Waves.** We shall restrict ourselves to the study of phenomena occurring when light is incident on the face of a crystal cut in two different ways, namely, perpendicular to the main axis and parallel to the main axis.

Light propagated along the main axis differs in no way from light propagated in isotropic bodies. The electric vector produces polarised oscillations of the

dipoles in the direction perpendicular to the main axis. Hence, the wave is propagated with the velocity  $v_0 = \frac{c}{n_0}$ , where  $n_0 = \sqrt{\epsilon_{\perp}}$ ;  $\epsilon_{\perp}$  is the dielectric constant for the direction perpendicular to the axis. The designations  $n_0$  and  $v_0$  indicate that the index of refraction and the velocity of light are "ordinary".

Recalling that  $\epsilon = 1 + 4\pi\alpha$ , we find that for Iceland spar,  $n_0 = \sqrt{1 + 4\pi\alpha_{\perp}} = 1.658$ . Hence,  $v_0 = 1.81 \times 10^{10}$  cm/sec.

The passage of light through such a crystal in the direction of the main axis does not change its polarised state. Natural light remains natural, and the oscillation direction of the electric vector for a polarised wave does not change.

The simplicity of the case considered is characteristic of a uniaxial crystal. Here, any polarised state of an incident wave is capable of exciting oscillations in the direction perpendicular to the main axis. Hence, the polarisability of the molecules (also  $\epsilon$  and  $n$ ) is the same for all such oscillations.

Now, let us consider the case of normal beam incidence on the face parallel to the main axis.

Different polarised waves behave differently. Consider the behaviour of a linearly polarised beam. If the electric vector is perpendicular to the axis, the light is propagated with the same velocity  $v_0$  as in the preceding case. But if the electric vector is parallel to the axis, the polarisation of the dipoles occurs along an axis for which the dielectric constant has another value, namely,  $\epsilon_{\parallel}$ . Therefore, for this propagation direction, the velocity and the index of refraction have

other values, namely,  $v_e = \frac{c}{n_e}$  and  $n_e = \sqrt{\epsilon_{\parallel}}$ , respectively. The designations  $n_e$  and  $v_e$  indicate that the index of refraction and the velocity of light are *extraordinary*. The reasons for the above designations will become evident below.

Crystals for which  $v_e < v_0$  are called optically positive; on the other hand, those for which  $v_e > v_0$  are called optically negative.

For Iceland spar,  $n_e = \sqrt{1 + 4\pi\alpha_{\parallel}} = 1.486$  and  $v_e = 2.02 \times 10^{10}$  cm/s. Iceland spar is an optically negative crystal since  $v_e > v_0$ .

**Elliptical Polarisation.** What is the situation when the electric vector  $E$  of the wave incident on the face of the crystal forms an angle  $\varphi$  with the direction of the main axis (Fig. 171)? Experiments show, and this may be predicted from Maxwell's equations, that the electromagnetic wave becomes divided into two parts. The vector  $E$  must be resolved into the components  $E \sin \varphi$  and  $E \cos \varphi$ . The first corresponds to a wave travelling with the velocity  $v_0$ , and the second to a wave travelling with the velocity  $v_e$ . This may be shown by determining the path difference between the two waves that are created upon division of the incident wave. Designating the thickness of the crystal by  $l$ , the phase difference can be expressed as  $\delta = \frac{2\pi}{\lambda} l (n_e - n_0)$ .

Thus, the polarised state of the wave leaving the crystal has changed significantly. The wave incident on the crystal was linearly polarised, while the one leaving is a combination of two waves having mutually perpendicular oscillation directions

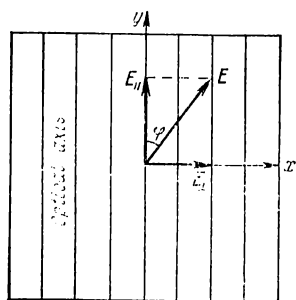


Fig. 171



and displaced relative to each other by  $\delta$ . What is the nature of this peculiar polarised state? Such light is said to be *elliptically polarised* since the terminus of the electric vector describes an elliptical spiral. If the electric vector of one of the waves is given by

$$E_x = E \sin \varphi \cos \omega t,$$

then for the other wave the electromagnetic oscillation in the plane perpendicular to the beam will have the form

$$E_y = E \cos \varphi \cos (\omega t + \delta).$$

The addition of such oscillations has been considered earlier (see p. 84). It was seen that the point describing two such oscillations is an ellipse. The same applies to the terminus of the electric vector, but, since the wave is advancing, the terminus of the vector  $E$  describes an ellipse in its projection on the plane perpendicular to the beam. In space, the terminus of vector  $E$  describes an elliptical spiral winding about the beam axis.

To obtain circularly polarised light by this method, a "quarterwave plate" is used. This is a plate producing a path difference of  $\lambda/4$  between waves travelling with velocities  $v_o$  and  $v_e$ . The thickness of such a plate must satisfy the equation

$$\frac{2\pi}{\lambda} l (n_o - n_e) = \frac{\pi}{2} + m\pi.$$

If a linearly polarised beam impinges on such a plate so that the vector  $E$  forms a  $45^\circ$  angle with the direction of the main axis of the crystal, resolution of this vector yields

$$E_x = \frac{E}{\sqrt{2}} \cos \omega t \quad \text{and} \quad E_y = \frac{E}{\sqrt{2}} \sin \omega t,$$

i.e.,

$$E_x^2 + E_y^2 = \frac{1}{2} E^2.$$

But this is the equation of a circle. Thus, the described experimental conditions lead to the transformation of linearly polarised light into circularly polarised light.

**Double Refraction.** The apparent bifurcation of objects viewed through a transparent crystal is a phenomenon that has been well known for a long time. It shows that division into two waves may occur not only as regards the propagation velocities, but also as regards the beam directions in space. Double refraction occurs for normal light incidence on a crystal face (ground or naturally formed), which is at an angle to the optic axis. The phenomenon may also be investigated by means of a plate cut parallel to the axis. In this case, the light must be incident at an angle to the normal. We shall direct our attention to the latter case. Let us introduce another restriction, namely, that the beam be directed in such a manner that the plane of incidence of the light is perpendicular to the optic axis.

Assume that a polarised beam impinges on the plate at an angle  $i$ . By turning the beam about its axis, the position of the electric vector relative to the plane of incidence is changed.

When the electric vector coincides with the incident plane (see Fig. 172a), no special effects are noted. Refraction occurs in accordance with the law for isotropic bodies, namely,  $\frac{\sin i}{\sin r_o} = n_o$ .

The refractive index turns out to be  $n_o$ . This is as it should be, for the electric vector is perpendicular to the main axis of the crystal. When the beam is turned

90° about its axis (see Fig. 172*b*), it is also refracted. But now  $\frac{\sin i}{\sin r_e} = n_e$ , i.e., the refraction angle is different and the index of refraction is that for an extraordinary beam. This too is natural, for the vector  $E$  coincides with the direction of the main axis.

Most remarkable is that an intermediate position does not yield a beam with an intermediate angle of refraction, but yields rather two beams—an ordinary and an extraordinary beam having refractive indexes  $n_o$  and  $n_e$ , respectively. As before, the field intensity vector is resolved into two vectors, one lying along the main axis and the other perpendicular to it. Each component creates its own field, or wave. In turning the beam of light about its axis, the intensities of these two beams continuously change; when one beam decreases in intensity, the other increases.

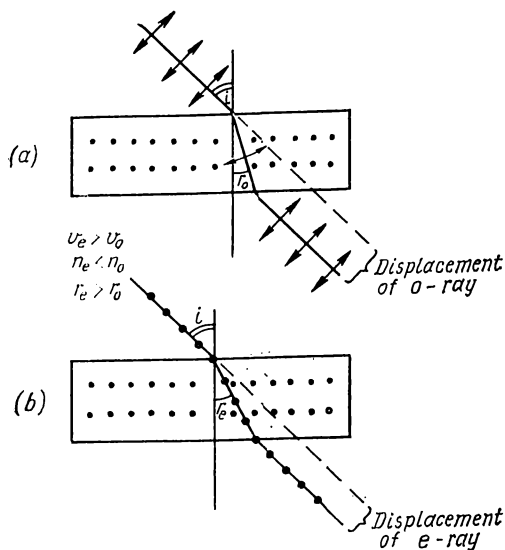


Fig. 172

Since the beams are refracted twice, i.e., in entering and leaving the plate, the ordinary and extraordinary beams emerge parallel to each other. The thicker the plate, the greater the separation between beams. If a narrow beam of incident light is used, the difference between the refractive indexes may be determined by measuring the beam displacements.

Now we can explain why the beams are called "ordinary" and "extraordinary". Let us begin turning a crystal plate, whose optic axis is parallel to the face, about the normal to the reflecting face. If we were dealing with an isotropic body, such rotation could not affect reflection and refraction. When we rotate the crystal plate

as indicated, nothing happens to one beam, i.e., its position in space and its intensity remain unchanged. That is how an ordinary beam behaves. It is understandable, therefore, that the beam whose electric vector is perpendicular to the main axis of the crystal is called ordinary. In this experiment, the electric vector component lying in the plane of incidence is always perpendicular to the main axis of the crystal. This component acts in an "ordinary" manner. On the other hand, the component of  $E$  perpendicular to the plane of incidence forms an angle with the main axis of the crystal that varies as the crystal is rotated. During such rotation, not only does the extraordinary beam's intensity vary, but its position in space varies as well. We see that the extraordinary beam does not obey the laws pertaining to isotropic bodies. In the general case, the refracted beam is not in the plane of incidence.

We shall not go into the rather complex explanation for these phenomena. It should be noted however, that these phenomena are in complete accord with Maxwell's electromagnetic field theory.

## Sec. 150. POLARISERS. INVESTIGATION OF THE POLARISED STATE OF LIGHT

It was stated on p. 253 that a dielectric placed with its plane surface at an angle  $\varphi_B$  to the incident beam may serve as a polariser. In this case, the reflected beam is completely polarised and the refracted beam is polarised to the maximum extent possible. However, it is not convenient to use a reflecting plate as a polariser, for the polarised beam travels at an angle to the incident beam. A stack of glass plates

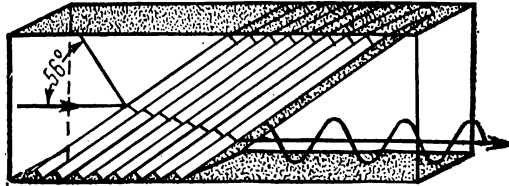


Fig. 173

may be used instead. By repeated refraction, an almost completely polarised beam may be obtained. However, a considerable portion of the light is absorbed in such a device (see Fig. 173).

The best kind of polariser is a crystal in which the linearly polarised ordinary (or extraordinary) beam may be separated out by means of double refraction. Such polarisers are known as Nicol prisms, or simply Nicols.

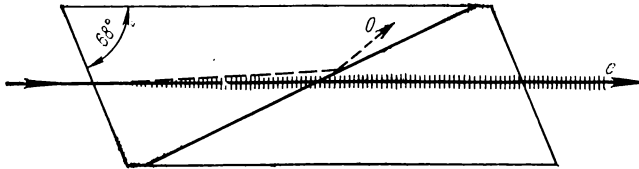


Fig. 174

The polariser proposed by the French scientist Nicol consists of two right-angled prisms made of Iceland spar (Fig. 174). These prisms are glued together with Canada balsam, a substance whose refractive index  $n$  is 1.550. This value lies between  $n_o$  and  $n_e$  for Iceland spar. A nonpolarised beam of light impinging on the prism is divided into two components. The ordinary beam is reflected at the boundary between the prisms, where the condition for its total reflection is satisfied, and is deflected to one side. The extraordinary beam passes through both prisms. Thus, a Nicol acts as a slit passing only electric vector oscillations directed in a specific direction.

Pleochroism is of great practical importance in polariser applications. The term is used to indicate that the ordinary and extraordinary beams are absorbed differently and that the absorption coefficient of the extraordinary beam is a function of the direction relative to the optic axis. A pleochromatic crystal gives a different hue and absorbs light differently when it is turned relative to the beam.

Tourmaline is a classical example of a pleochromatic crystal. The absorption coefficient for the ordinary beam over almost the entire visible spectrum is so great that a tourmaline plate of 1 mm thickness, cut parallel to the optic axis, transmits in effect only the extraordinary beam and, hence, may serve as a polariser. How-

ever, the yellowish-green hue of the transmitted light prevents tourmaline from being used in practice as a polariser.

*Polaroids*, synthetic pleochromatic films, have wide application and may be prepared from herapathite (quinine sulphate periodide), a strongly pleochromatic substance. A polaroid is a transparent plastic film consisting of submicroscopic crystalline herapathite needles oriented in a single direction. To orient the crystals,

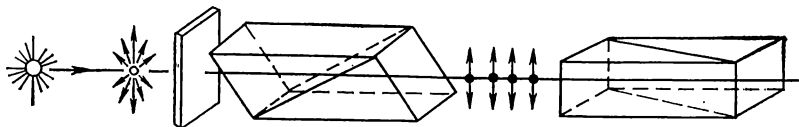


Fig. 175

a viscous mass of the crystals is placed between two glass plates and subjected to reciprocating motion. To be sure, herapathite is not the only substance suitable for the production of polaroids.

Iodine atoms, constituents of herapathite, are essential to secure pleochromatic properties. Pure iodine polaroids may be produced by iodising thin films of polyvinyl alcohol.

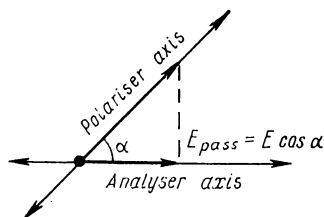


Fig. 176

The polarised state of light may be investigated with the aid of two Nicol prisms, or other polarisers, and a "quarter-wave plate". A 0.038 mm-thick sheet of mica may serve as the quarter-wave plate. Let us consider the passage of light through two Nicol prisms. In order to distinguish between the two, the first prism in the path of the beam is called the polariser and the second the analyser (see Fig. 175). If a beam of natural light of intensity  $I_0$  impinges on the polariser, a linearly polar-

ised light of intensity  $\frac{1}{2} I_0$  emerges from the prism. Naturally, turning the polariser about its axis in no way changes the intensity of the transmitted beam. By means of the analyser, it may be shown that the Nicol prism has indeed transmitted a linearly polarised beam. If the Nicol prisms are oriented with their "slits" parallel to each other, the light will be transmitted through the analyser as well without change in intensity (disregarding absorption in the material of the polarising device). For crossed prisms, i.e., when the "slits" are at right angles to each other, light is not transmitted (see Fig. 175). The intensity of the light when an angle  $\alpha$  is formed between "slits" is  $I = \frac{1}{2} I_0 \cos^2 \alpha$ . Thus, the electric vector of the wave arriving at the analyser may be resolved into two components—one parallel to the "slit" and the other perpendicular to it. Since the component that is passed is  $E \cos \alpha$  (see Fig. 176), the intensity is proportional to  $\cos^2 \alpha$ .

The first prism will already reveal whether or not the light was partially or completely polarised.

Using two Nicol prisms, it is not possible to distinguish circularly polarised light from natural light, and elliptically polarised light from partially polarised natural light. To do this, the quarter-wave plate may be used. If it is placed before the polariser, it in no way affects naturally polarised light, but circularly polarised light is transformed into linearly polarised light. Similarly, a quarter-wave plate changes the properties of elliptically polarised light.

## Sec. 151. A CRYSTAL PLATE BETWEEN "CROSSED" NICOL PRISMS

The method of investigating transparent anisotropic substances by observing the behaviour of linearly polarised light impinging on them is very widespread. In order not to complicate the problem, assume that we are dealing with a crystal plate cut parallel to the optic axis. This plate is placed between Nicol prisms.

From the polariser, a linearly polarised beam impinges on the plate. Remove the plate and place the analyser in a crossed position. Light is not transmitted. Now put the plate back. The field becomes illuminated, i.e., the beam of light is transmitted through the system. There can be only one explanation, namely, the crystal plate has changed the polarised state of the beam coming from the polariser. By means of the analyser, we may determine the exact nature of the change. If upon turning the analyser a new position for darkness is found, the conclusion is that the crystal plate has changed the direction of the beam oscillations but has left them linearly polarised. If the intensity of the light is not changed by turning the analyser, the conclusion is that the plate has transformed the linearly polarised light into circularly polarised light. Finally if the light is not extinguished by turning the analyser or plate, but the intensity of the light is changed thereby, the conclusion is that the plate has created elliptically polarised light.

The changes produced in linearly polarised light depend on two things—the mutual orientation of the plate's optic axis and the oscillation direction of the beam coming from the polariser, and the phase difference  $\delta = \frac{-\pi}{\lambda} (n_o - n_e)$  created by the plate between the ordinary and extraordinary waves into which the incident wave is divided.

If the substance placed between the crossed Nicol prisms is isotropic, no illumination of the field occurs. The phenomenon described above may be utilised in the investigation of anisotropic substances.

Usually, observations are made using crossed Nicol prisms between which the plate is rotated. During such rotation, the illumination does not remain constant. At every instant, the amplitude  $A$  of the light emerging from the polariser is resolved into the components  $A \cos \varphi$  and  $A \sin \varphi$ , where  $\varphi$  is the angle between the polariser "slit" and the optic axis of the plate. The terminus of the electric vector of the wave emerging from the plate describes an ellipse:

$$\frac{E_x^2}{A^2 \cos^2 \varphi} + \frac{E_y^2}{A^2 \sin^2 \varphi} - \frac{2E_x E_y}{A^2 \cos \varphi \sin \varphi} \cos \delta = \sin^2 \delta,$$

where  $\delta$  is fixed and  $\varphi$  changes continuously. Fig. 177 shows ellipse transformations for the case of a quarter-wave plate, i.e., in the above formula  $\delta = 90^\circ$ . For different values of  $\varphi$ , different polarised states are obtained.

Since the path difference  $\delta$  depends on the wavelength, the pattern is coloured when white light is used. If the plate is of uniform thickness, it will have a single colour that differs for every different relative orientation of the plate and Nicol prisms. Thus, for certain wavelengths of white light, the plate thickness may be equal to  $\frac{\lambda}{4}$ , for others it may be equal to  $\frac{\lambda}{2}$ , and for still others a multiple of  $\lambda$ .

Accordingly, for different wavelengths there arise different polarised states, which are transmitted by the analyser to a greater or lesser degree for different relative orientations of the plate and Nicol prisms. The phenomenon of chromatic polarisation is very beautiful. It is hardly likely that the richness of shades and hues observed when the thickness of a plate changes (e.g., in crystal growth), or when the

orientation of a plate is varied relative to Nicol prisms, may be achieved by any other means.

If a plate is of variable thickness, the interference fringes are rainbowed when observed in white light.

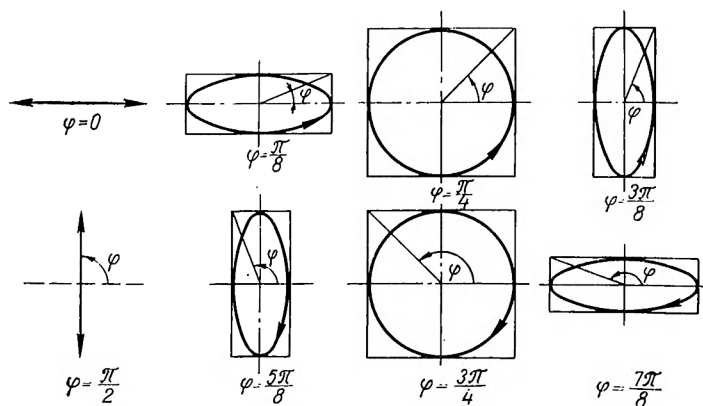


Fig. 177

In addition to these patterns of fringes representing equal thickness, distinctive fringes representing equal inclination may be observed if the crystal plate is observed using converging rays (iconometric investigation). These observations may be made on small crystalline granules in the field of vision of a microscope. Their practical significance lies in the determination of crystal symmetry. In particular, it is not difficult to determine to which of the following three groups an object belongs: (1) amorphous or crystalline substances having cubic symmetry; (2) uniaxial crystals; and (3) crystals having symmetry of lower order. In the first case, interference fringes are not present. Uniaxial crystals give the pattern shown in Fig. 178 when the optic axis coincides with that of the incident beam of light. A dark washed-out cross is formed in the area where either an ordinary or an extraordinary beam impinges. If elliptically polarised oscillations are not transformed, crossed Nicol prisms do not transmit light.

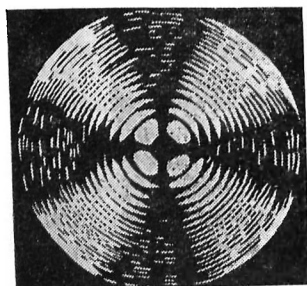


Fig. 178

The theory of this phenomenon is too extensive to be dealt with here.

## Sec. 152. DOUBLE REFRACTION DUE TO AN EXTERNAL ACTION

As was stated at the beginning of this chapter, almost every molecule possesses anisotropic polarisability. An optically isotropic body may be formed from anisotropic molecules if the latter are randomly distributed. On the other hand, if a preferred orientation is given to the molecules by any means, the polarisation vector, i.e., the total dipole moment of the molecules, no longer has the same value in every direction, the body acquires permittivity anisotropy and hence optical anisotropy.

Such anisotropy, expressed in the manifestation of double refraction, is almost always produced in nonuniformly deformed solid bodies, certain liquids placed in electric fields (Kerr effect), and streams of liquids whose molecules are of elongated form. Double refraction may be observed in biological objects and high-polymers, i.e., again in substances consisting of long molecules which cannot lie in a completely random manner in the substance. Generally speaking, to one or another degree double refraction is almost always present since it is very difficult to make a body ideally isotropic.

If a body is subjected to one-sided compression or extension, an axial type of anisotropy is produced. With respect to optical properties, such a body is similar to a uniaxial crystal. Optical anisotropy is most conveniently observed between crossed Nicol prisms. Lightly pressing a transparent piece of plastic or glass between one's fingers makes it anisotropic and, as a result, the field of vision is immediately illuminated. At different locations in the object, a nonuniform deformation produces different values for the difference  $n_o - n_e$ . Therefore, fringes

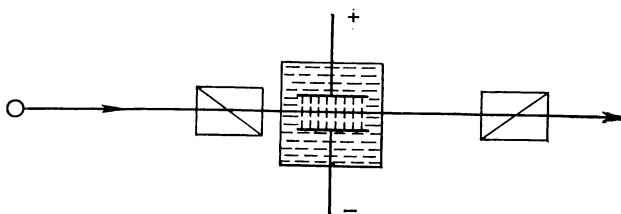


Fig. 179

representing equal phase difference  $\delta$  are formed on a body subjected to deformation. The forms of these curves correspond to the stresses produced in the body. By analysing the nature of these curves, it is possible to obtain a clear picture of the stress distribution.

Since this valuable method is applicable only to transparent bodies, how can it be used in practice? The approach is clear: construct transparent plastic models. By constructing a model of a bridge, building or machine element and loading it proportionately, a picture of the produced stress may be obtained. The optical method of quantitatively determining stress is known as the photoelastic method, which is applied under the guidance of specialists in this field.

Queer coloured patterns appear when observed in white light. If a deformation is elastic, the pattern vanishes upon removal of the load. On the other hand, if the stress remains upon removal of the load, the coloured pattern also remains.

Now, let us consider an optically anisotropic liquid located in an electric field. The electric field exerts an orienting action only if the molecules of the liquid possess a constant, rigid dipole moment.

In this case, the molecules tend to become oriented in such a manner that the direction of the rigid dipole moment coincides with the direction of the field. Thus, the liquid acquires the properties of a uniaxial crystal whose optic axis is parallel to  $E$ . This effect is best observed when the liquid and the applied field are located between crossed Nicol prisms (see Fig. 179).

Experiments show that the difference  $n_o - n_e$  produced in the liquid is proportional to the electric field intensity  $E$  squared. The phase difference is given by  $\delta = BLE^2$ , where  $l$  is the distance traversed in the liquid by the beam of light and  $B$  is the Kerr constant (characteristic of the substance).

Nitrobenzene, whose rigid dipole moment is large, is notable for its large value of  $B$ :  $2 \times 10^{-5}$  (CGS units). Benzene has a Kerr constant of  $0.5 \times 10^{-7}$  and carbon disulphide  $3.5 \times 10^{-7}$  (CGS units).

*Example.* A 10 cm-long condenser filled with nitrobenzene will act as a quarter-wave plate if the field intensity in it is  $E = \sqrt{\frac{\pi/2}{B\ell}} = 26,600$  V/cm. For this purpose, a potential difference of 2,660 V must be applied to the condenser if the distance between plates is 1 mm.

The Kerr effect makes it possible to transform electric field oscillations into light intensity variations. This effect has little inertia: the relaxation time, i.e., the time required for the molecules to assume the appropriate orientation in the electric field, is of the order of one-thousand millionth of a second. Therefore, electric oscillations modulated by sound may be transformed into light intensity variations. This makes it possible for sound to be recorded on photographic film.

### Sec. 153. OPTICAL ACTIVITY

The ability of certain substances to change the oscillation direction of a linearly polarised beam is known as optical activity. The phenomenon may be described as follows. Consider an arrangement of crossed Nicol prisms with an optically active substance placed in the path of a beam. The field becomes illuminated, but the illumination disappears when the analyser is turned by some angle  $\alpha$ . Thus, linearly polarised light transmitted through an optically active substance remains linearly polarised, but the oscillation direction of the beam changes by the angle  $\alpha$ . Experiment shows that the change in oscillation direction is strictly proportional to the thickness of the layer of substance:

$$\alpha = \rho d$$

The constant  $\rho$  characterising the substance is known as the specific rotation constant and is usually expressed in degrees per millimetre. The phenomenon exhibits dispersion since  $\rho$  is a function of the wavelength.

The change in oscillation direction is quite considerable and in the case of many substances attains a value considerably greater than ten degrees per mm for a number of wavelengths. For water solutions of organic substances, the rotation of the polarisation plane is a function of the concentration, i.e.,  $\alpha = \rho cd$ , where  $c$  is the concentration.

What kind of substances are optically active? An optically active substance must be composed of structural units which have neither a plane of symmetry nor a centre of symmetry among their elements of symmetry. In the case of molecular substances, such components are, as a rule, molecules. In the case of crystals, in which molecules are not distinguishable, such components are unit cells.

Molecules, or cells, satisfying the above conditions, may be encountered in the form of two optical isomers, designated by the letters  $d$  and  $l$  (dextro and levo, or right and left). An object and its image in a mirror are optically isomeric. A substance consisting of  $d$ -molecules (or cells) rotates light to the right, while one consisting of  $l$ -molecules rotates light to the left. By rotation to the right we mean the case in which the analyser must be turned to the right (from the viewpoint of an observer facing the oncoming beam of light) to restore darkness when the thickness of the layer of substance is increased. Reversing the direction of the light does not change the sign of the effect.



Optical activity may occur for substances in the liquid as well as in the solid state. It is merely necessary that there be a surplus of *d*- or *l*-molecules. The orientation of these molecules may be either random or uniform. In the first case, the body is isotropic and the rotation is the same irrespective of the direction of the beam of light. In optically active crystals, the magnitude of the rotation  $\alpha$  is a function of the beam direction relative to the crystal axes.

When molecular crystals that rotate light are fused, the structural components remain intact. In such cases, the solid as well as liquid substance possesses optical activity. An example of this is sugar, which, in addition, possesses activity in solution. This property of sugar is utilised in the saccharimeter to determine the amount of sugar in a solution from the magnitude of the change in the oscillation direction of a beam of light.

The situation is different in such crystals as quartz (see Fig. 180). The arrangement of atoms in a quartz cell satisfies the necessary conditions, namely, it does not have a centre of symmetry nor a plane of symmetry. The molecules are not distinguishable in a quartz crystal; as a result, the arrangement of atoms changes upon fusion. Hence, in fused quartz the necessary structural units are not present. Fused quartz is, therefore, optically inactive.

One and the same substance, from the point of view of chemical composition, may be encountered in the optically active form and in the optically inactive form. This applies not only to quartz. The structure of an inactive variety bears little resemblance to the structure of crystals possessing optical activity.

The above is quite understandable in the case of ionic and homopolar crystals. But how can an inactive crystal be formed from active molecules in the case of a molecular crystal? This occurs by the formation of racemic crystals. A racemic mixture is a mixture of equal quantities of *d*- and *l*-molecules. Such a mixture does not rotate light since the two opposite effects are equalised. A racemic crystal consists of pairs of *d*- and *l*-molecules. Every pair constitutes a centrosymmetric group of atoms.

Optically active crystals exist in both the *d*- and *l*-forms. The structures of such crystals are identical in the same sense that right and left gloves are identical. In the case of molecular crystals, this means that in one case the structure is composed of *d*-molecules and in the other of *l*-molecules. Dextro- and levo-quartz, dextro- and levo-glucose, dextro- and levo-tartaric acid—all the properties of these substances, all the details of their structure, are the same except for the fact that light is rotated in different directions.

Inorganic optical isomers (e.g., dextro- and levo-quartz) are encountered in nature in equal quantities. This is not the case with organic molecules, which are important in biology. The French chemist Pasteur showed that a number of micro-organisms are capable of feeding on only a specific optical isomer.

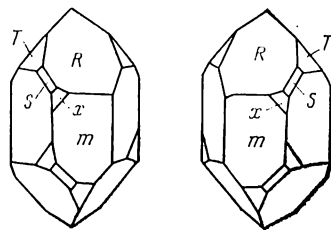


Fig. 180

## Sec. 154. BASIC THEORY OF OPTICAL ACTIVITY

How is the phenomenon of optical activity explained? Before answering this question, we shall show that a linearly polarised beam is equivalent to two circularly polarised beams that are dextro- and levo-rotated.

Let us write the equations for the electric vector oscillations, taking into account that there is a phase difference  $\delta$  between these circularly polarised waves. For the dextro-rotated wave

$$E_x^d = E_0 \cos \omega t \quad \text{and} \quad E_y^d = E_0 \sin \omega t;$$

for the levo-rotated wave

$$E_x^l = E_0 \cos(\omega t + \delta) \quad \text{and} \quad E_y^l = -E_0 \sin(\omega t + \delta).$$

The total field has the components

$$E_x = E_x^d + E_x^l \quad \text{and} \quad E_y = E_y^d + E_y^l.$$

To determine the polarised state of the resulting oscillation, let us calculate the ratio  $\frac{E_y}{E_x}$  for the total field. Using simple trigonometric transformations, we obtain

$$\frac{E_y}{E_x} = -\tan \frac{\delta}{2}.$$

The ratio is independent of time. It is seen that we are dealing with linearly polarised oscillations that form an angle  $\frac{\delta}{2}$  with the  $x$ -axis. Q.E.D.

On the basis of this conception, the phenomenon of rotation of oscillation direction is quite easily understood. Rotation of the oscillation plane by an angle  $\frac{\delta}{2}$  signifies that the levo-rotated wave is lagging behind the dextro-rotated wave (or vice versa, depending on the rotation direction) by the angle  $\delta$ . In view of this explanation, it is clear why a discussion of optical activity has been included in the chapter on double refraction. Here, too, the wave is divided by the substance into two components, one moving faster than the other and continuously advancing relative to it in phase. From this viewpoint, the specific rotation is proportional to the difference in refractive index between the dextro- and levo-rotated beams.

This discussion has in no way advanced our understanding of the phenomenon. We have merely given it another (completely equivalent) interpretation. However, this new approach enables us to more easily explain optical activity. Circularly polarised waves that are dextro- and levo-rotated travel through a substance with different velocities. They encounter different indexes of refraction and, hence, different permittivities and polarisabilities. The displacements of the electron cloud under the action of these two waves must differ. One wave experiences more difficulty than the other in displacing electrons from their equilibrium positions. If we can determine the reason for this difference, the explanation for optical activity will have been found.

We know from chemistry that if a molecule has an asymmetrical carbon atom, the substance may exhibit optical activity. By an asymmetrical carbon atom chemists mean a carbon atom bound to each of four different atoms or radicals.

The angles formed between the bonds of a tetravalent carbon atom are approximately equal to those in a tetrahedron. Fig. 181 shows a molecule containing an asymmetrical carbon atom. The radicals or atoms bound to C differ, but their nature is unimportant. We see, in the first place, that two different molecules of such a substance, which are mirror images of each other, are possible. These are optical isomers and cannot be made to coincide. This may be easily demonstrated using wire models.

Consider a circularly polarised wave travelling along the axis of symmetry of the bonds. In Fig. 182 the wave is directed outward from the page. Atoms  $A$  and  $B$  are higher than atoms  $D$  and  $E$ . Let us determine the directions of the electron displacements for dextro- and levo-rotated waves. Assume the situation is as follows for the dextro-rotated wave: when the vector  $E$  is directed along  $ED$ , in the upper "level" it is directed along  $BA$ . If that is the case, the situation is as follows for the levo-rotated wave: when the vector  $E$  is directed along  $ED$ , in the upper "level" it is directed along  $AB$ .

Examining the figures, we see that the displaced electrons behave differently. In the first case, the electrons of atoms  $A$  and  $D$  move simultaneously away from

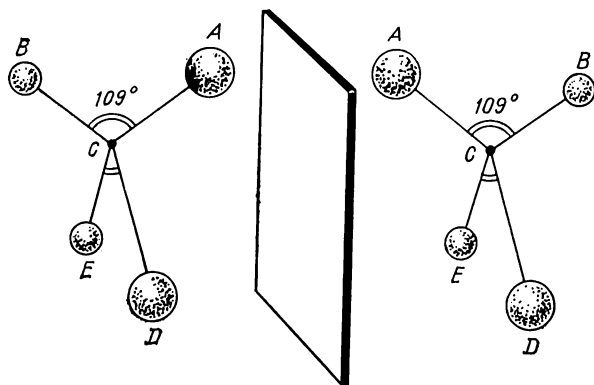


Fig. 181

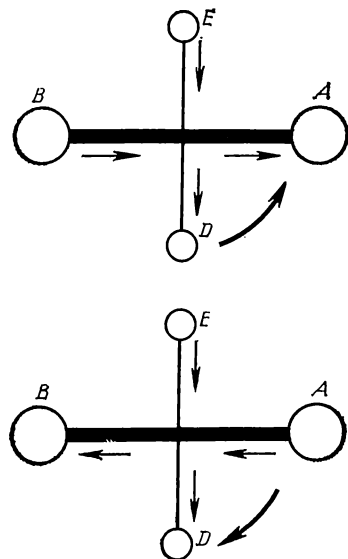


Fig. 182

the centre. In the second case, when the electrons of  $A$  move toward the centre, the electrons of  $D$  move away from the centre. Such differences will always be found for systems of atoms that do not have a centre of symmetry and a plane of symmetry. On the other hand, if these elements of symmetry are present, the action of dextro- and levo-rotated waves will be the same.

Different electron displacement conditions for asymmetrical groupings lead to different polarisabilities and, hence, to different indexes of refraction for dextro- and levo-rotated waves. It is clear that the behaviour of a dextro-molecule is opposite to that of a levo-molecule.

A deeper theoretical analysis shows that the effect is maintained for all orientations of the molecules relative to the beam.

# The Theory of Relativity

## Sec. 155. BASIC THEORY

The theory of relativity, developed at the beginning of this century by the great modern physicist Albert Einstein, is based on two postulates: (1) the principle of relativity and (2) the principle of constancy of the velocity of light. We shall briefly consider the essence of these principles, describe the experiments confirming them, and discuss certain consequences of the theory.

The theory of relativity originated with the questioning of the existence of a mechanical carrier (ether) for an electromagnetic field. The theory of relativity solved this problem and in this sense may be viewed as the perfection of electromagnetic field theory. While solving the problems posed by electrodynamics, the theory of relativity went much further. Its development led to the establishment of the laws of mechanical motion at velocities close to that of light, to the law of the equivalence of mass and energy, and to new views on the nature of gravity. Since our discussion will be, of necessity, very brief, we are forced to dispense with an historical narration.

First, as to the essence of the main postulates. The principle of relativity states that all laws of nature (and not only the laws of mechanics) are the same in all inertial systems of coordinates. The principle states that not a single physical experiment could discover special properties for one of the inertial systems. All inertial systems are equivalent.

The second postulate pertains to the constancy of the velocity of light in a vacuum for all inertial systems. From this it follows that the velocity of light in the "receding" and "approaching" directions must be the same, i.e., that the velocity of light is independent of the light source and measuring instruments.

How do these principles affect our views concerning an electromagnetic field and its carrier? It is not difficult to see from the formulations of the principles that electromagnetic waves and, say, sound waves, are not analogous.

Imagine a laboratory isolated from the external world, moving rectilinearly and uniformly relative to the stars. In this laboratory, measurements are made of the velocity of sound in the direction of motion. Theoretically, two extreme cases are possible: in one, the walls of the laboratory are impervious to air, so that the air is carried along by the laboratory; in the other, the walls are pervious to air, the air is stationary relative to the stars, and the laboratory moves through the air, i.e., without carrying it along. Assume in these two cases that measurements are made of the velocity of sound. The velocities are measured by two observers—one moving and the other stationary relative to the stars. In each case, the velocities of sound relative to these two observers will differ. If the velocity of sound in air is designated by  $c$  and the velocity of the laboratory relative to the stationary observer by  $v$ , then, in the case in which the air is carried along, the moving observer finds the velocity equal to  $c$  and the stationary observer finds it equal to  $c + v$ ; in the case in which the air is not carried along, the moving observer finds the velocity equal to  $c - v$  and the stationary observer finds it equal to  $c$ .

The postulates of the theory of relativity reject both variants in the case of an electromagnetic wave in an ether. In experiments with light waves, the velocity of light will be equal to  $c$  for the stationary as well as the moving observer. This means that a stationary as well as a moving ether is incompatible with the theory

of relativity. Thus, the theory of relativity rejects the possibility of viewing the field as a medium in which mechanical displacements occur. We must conclude that electric and magnetic fields have a real existence.

#### Sec. 156. EXPERIMENTAL VERIFICATION OF THE PRINCIPLE OF CONSTANCY OF THE VELOCITY OF LIGHT

At first glance, the principle of constancy of the velocity of light seems to fly in the face of "common sense". Therefore, before discussing certain consequences of the theory of relativity, it is desirable to describe the direct experimental evidence for its validity. This evidence is derived from astronomical observations.

Astronomers have discovered the existence of so-called double stars. A double star consists of two heavenly bodies of approximately the same mass rotating about their overall centre of gravity. We have the means to measure the distance between the stars, their mass and their velocity; also, to determine their relative motion. If the velocity of light depended on the velocity of the star itself, the velocity of the heavenly body would be added to the velocity of light when this body moved toward a terrestrial observer and subtracted when this body moved away from the terrestrial observer. In such a case, to the terrestrial observer, the motion during one half of the orbit would appear faster than during the other half. This effect would be detectable even if the velocity  $v$  of the heavenly body were one-hundred thousandth of the velocity of light  $c$ .

Thus, for a long distance  $l$ , the difference in times  $\frac{l}{c-v}$  and  $\frac{l}{c+v}$  may be so considerable, even for very small values of  $v$ , that not only is the periodicity disturbed, but a light beam transmitted during motion in the "receding" direction may overtake a beam transmitted during motion in the "approaching" direction. Then, rotation of the star would not be visible or would be of a peculiar nature. The periodic rotation of double stars may be understood only on the basis of the principle of constancy of the velocity of light.

To be sure, our discussion has dealt with the motion of a light source, so that there may remain some doubt regarding the validity of the principle of constancy of velocity for the motion of an observer. Such doubt may be removed by another astronomical observation, i.e., observation of the periodicity of the motion of Jupiter's satellites. Measurements of the motion of Jupiter's satellites may be made in two cases—when the light arriving on the Earth from Jupiter coincides with the direction of motion of the solar system and when it is in the opposite direction. The identicalness of the observations and the distinct periodicity of Jupiter's annual motion demonstrate the validity of the principle of the constancy of the velocity of light in this case as well.

The most important role in the development of the theory of relativity was played by an experiment first performed by Michelson in 1881 with the aid of the interferometer described on page 272. This experiment consisted in the following. The locations of two mirrors, i.e., the arm lengths  $l_1$  and  $l_2$ , were selected in such a manner that the coherent beams into which a light signal is divided would require the same amount of time to cover the distances along the two arms of the interferometer. This selection is made when the interferometer is arranged in such a manner that one of the arms is parallel to the motion of the globe in its orbit. The instrument is then turned  $90^\circ$  and the interference fringes observed for possible displacement.

The results of the Michelson experiment, which was repeated many times by Michelson and other investigators, are the following: no displacements of the

fringes occur and the times required for light to cover the distances along the arms remain equal when the instrument is turned  $90^\circ$ . This conclusion is based on very accurate measurements.

What is the significance of this experiment? Since the Earth moves with a velocity  $v \approx 30$  km/sec relative to fixed stars, the distances covered by the two rays cannot be the same from the viewpoint of a celestial inertial observer.

Let us examine the paths of the two rays (see Fig. 183). Of course, we need only concern ourselves with the portions of the path along which the beams travel

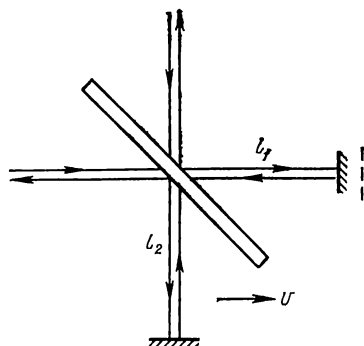


Fig. 183

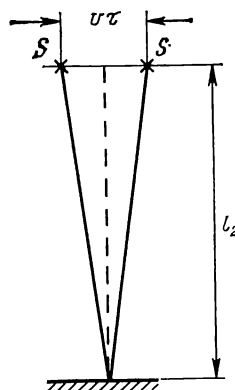


Fig. 184

separately. The longitudinal beam in the “receding” direction must cover the distance of the arm length  $l_1$  and overtake the mirror moving with a velocity  $v$  in the same direction. Therefore, the distance  $c\tau_1$  covered by the beam must equal  $l_1 + v\tau_1$ . The time required for the wave front to reach the mirror is

$$\tau_1 = \frac{l_1}{c \left(1 - \frac{v}{c}\right)}.$$

In the “approaching” direction, the beam covers the distance of the arm length  $l_1$  minus the distance covered by the approaching instrument. Therefore, the distance  $c\tau_2$  covered by the beam must equal  $l_1 - v\tau_2$ . Then,

$$\tau_2 = \frac{l_1}{c \left(1 + \frac{v}{c}\right)}.$$

The time  $\tau_1 + \tau_2$  measured in the experiment is equal to

$$\frac{2l_1}{c \left(1 - \frac{v^2}{c^2}\right)}.$$

Now, let us direct our attention to the transverse beam (see Fig. 184). During the total time  $\tau$ , i.e., the time elapsed from the instant the beam leaves the centre of the instrument  $S$  to the instant it returns, the mirror is displaced as shown in the figure. Therefore, the distance covered by the wave is

$$c\tau = 2 \sqrt{l_2^2 + \left(\frac{v\tau}{2}\right)^2};$$

whence, the time  $\tau$  is equal to

$$\frac{2l_2}{c \sqrt{1 - \frac{v^2}{c^2}}}.$$

In the first measurement, the arms  $l_1$  and  $l_2$  were selected in such a manner that the times required for the beams to cover the separate paths were equal. Hence,

$$\frac{2l_1}{c \left(1 - \frac{v^2}{c^2}\right)} = \frac{2l_2}{c \sqrt{1 - \frac{v^2}{c^2}}} \quad \text{or} \quad \frac{l_1}{\sqrt{1 - \frac{v^2}{c^2}}} = l_2.$$

However, in the second experiment, i.e., when the interferometer is turned  $90^\circ$ , there is no interference fringe displacement and the times remain equal even though the arms  $l_1$  and  $l_2$  have interchanged places! This is the surprising result of this experiment.

Thus, if the first arm is longitudinal,

$$l_1 = l_2 \sqrt{1 - \frac{v^2}{c^2}};$$

if the second arm is longitudinal,

$$l_2 = l_1 \sqrt{1 - \frac{v^2}{c^2}}.$$

For  $v = 0$ ,  $l_1 = l_2$ ; but for  $v \neq 0$ , we obtain a remarkable result: the length of one and the same segment differs, depending on whether this segment is parallel to the direction of motion or perpendicular to it. The obtained result is valid for any body and for any distance between two points. Thus, the first consequence of the theory of relativity is that a body moving relative to an inertial observer shortens its dimension in the direction of motion. The transverse dimensions remain unchanged. If an observer relative to whom an object is stationary finds that the length of this object is  $l_0$ , an observer relative to whom this object is moving with velocity  $v$  will find that its length is

$$l = l_0 \sqrt{1 - \frac{v^2}{c^2}}.$$

*Example.* If an object moves with a velocity of 1,000 km/sec relative to some "stationary" observer, the length of the object in the direction of motion appears to be  $l$  divided by 1.000005. If the velocity of the object is 200,000 km/sec, then  $\frac{l_0}{l} = 1.34$ .

The length of one and the same segment moving in a specific manner in different frames of reference, will differ. It is necessary to understand properly the relative nature of the above contraction. Take two rods of the same length  $l_0$  and assume that their relative velocity is  $v$ . Now, assume that there are two observers—one moving with the first rod and the other with the second. In such a case, the first observer will find that his rod has a length  $l_0$  and the other a length  $l_0 \sqrt{1 - \frac{v^2}{c^2}}$ . For this observer, the second rod will be shorter than the first. On the other hand, the second observer finds that the second rod has a length  $l_0$  and the first a length  $l_0 \sqrt{1 - \frac{v^2}{c^2}}$ . For him, the first rod is shorter than the second.

The length of a rod (or, in general, the distance between two points) is a relative concept. Of all rod lengths measured in various inertial systems, the rest length  $l_0$  is outstanding. This maximum length of a rod has absolute meaning.

## Sec. 157. TIME IN THE THEORY OF RELATIVITY

In the expression relating the length of a rod at rest to the length of a rod in motion, the factor  $\sqrt{1 - \beta^2}$  appears, where  $\beta = \frac{v}{c}$ . This factor also appears in analogous formulas relating the values of various physical quantities for stationary and moving observers. Using an approach similar to that taken in the preceding article leads to interesting results as regards time and acceleration, mass and force, momentum and energy, density of charge and current, field intensities, etc. The formulas of the theory of relativity enable us to convert values determined by a stationary observer to values determined by a moving observer. The ratio  $\beta = \frac{v}{c}$  is in all cases an important criterion of the need for a relativistic correction.

It is easily seen that  $\beta^2$  is comparable to unity only when the velocity is very large. Even when  $v = 100,000$  km/s,  $\sqrt{1 - \beta^2}$  is only several per cent less than unity. It is, therefore, clear that the theory of relativity yields negligible corrections when the motion is slow, i.e., in such cases it is not necessary to take into account the changes in physical properties with motion. The theory of relativity is of particular importance for the microworld, where particles having velocities approaching the velocity of light are encountered quite often.

Let us direct our attention to the consequences of the theory as regards time. It turns out that the interval  $\tau$  during which an event occurs is also not the same from the viewpoint of two different inertial systems. Thus, two events occurring simultaneously from one viewpoint occur at different times—one earlier and the other later—from the viewpoint of another frame of reference.

Qualitatively, this assertion follows immediately from the principle of the constancy of the velocity of light. Thus, consider a system moving uniformly and rectilinearly relative to another inertial system. In one of them, there is located a radiator from which light is radiated in all directions. In this system, let us select two points equidistant from the source of light along a straight line in the direction of relative motion. It is clear that in this system the light arrives at both points simultaneously. This is the situation from the viewpoint of an observer moving together with the source. However, to an observer in the other system the situation appears to be different. To this observer, one point is moving toward the signal and the other away. Since the velocity  $c$  has the same value for this observer too (moreover, the same in both directions), from his viewpoint the light arrives earlier at the point that is behind.

A doubt may arise: cannot such a conclusion lead to absurdities? One may reason that since the concept of simultaneity is relative, it may happen that from the viewpoint of one frame of reference a gun is fired and then a wounded bird falls from a tree, but from the viewpoint of another frame of reference the bird falls before the gun is fired. Careful analysis shows that the relativity of a sequence of events is limited by the velocity of propagation of interaction (less than  $c$ ). Therefore, "earlier" and "later" may interchange places only when they are not causally related, i.e., when they are not the result of interaction.

A very interesting result of the theory relates to the proper time of an object, i.e., the time determined by a clock moving together with a given body. If a time  $\tau$  has elapsed according to the clock of an observer in a certain inertial system, the handle of the clock moving with the object will have advanced by the time

$$\tau_0 = \tau \sqrt{1 - \beta^2}.$$



This means that a clock moving in any arbitrary manner moves slower than a stationary clock.\*

It is necessary to comprehend properly the relative meaning of this assertion. If two observers are in different inertial systems, each will assert that the clock of the other observer is slow. This would seem to be a paradox. Let us stop the observers and compare their clocks. However, to perform this check, at least one of the observers must perform a complete circuit. Upon returning to the point of departure, the clocks may be compared. But now the determinations of the observers have lost their relativity. The observer remaining in an inertial system is justified in applying the above formula. The observer executing the circuit has undergone accelerated motion; hence, the formula  $\tau_0 = \tau \sqrt{1 - \beta^2}$  cannot be used by him. Thus, after the observer executing the circuit returns and the clocks are compared, it turns out that his clock is slow. Moreover, he cannot "dispute" this result by reference to the principle of relativity, for this principle is valid only for inertial observers.

### Sec. 158. MASS

If the mass of a body measured in a system of coordinates to which it is bound is designated by  $m_0$ , to an observer relative to whom this body moves the mass appears to be

$$m = \frac{m_0}{\sqrt{1 - \beta^2}}.$$

The quantity  $m_0$  is known as *the rest mass* and the increase in mass with increasing velocity is a natural consequence of the fundamental principles of the theory. The velocity of light  $c$  constitutes a limiting velocity for any motion or transfer of interaction. For  $v = c$ , the mass of a body becomes infinite. Of course, the closer a body approaches the limiting velocity, the more difficult it is to accelerate it.

The increase in mass with increasing velocity was first detected for the electrons of  $\beta$ -rays as early as the beginning of this century. Since the electron velocity  $v$  and the ratio  $\frac{e}{m}$  may be determined independently (see p. 351), and since the electron charge remains unchanged, we are able to check the formula for mass.

Corrections given by the factor  $\sqrt{1 - \beta^2}$  play an important role in the construction of accelerators of charged particles. The particle velocities attained in modern accelerators are so great that, for example, in one of them  $\beta$  reaches a value of 0.9986; thus, the mass becomes 60 times heavier than the rest mass. In all experiments conducted under terrestrial conditions with macroscopic bodies, we can disregard the  $\sqrt{1 - \beta^2}$  correction to the value of the mass. Nevertheless, it is desirable to check its validity not only for elementary particles. This is possible by means of precise astronomical observations. It turns out that the change in mass of the planet Mercury during its orbital motion explains the small deviations of the orbit from an ellipse.

The momentum formula acquires the following form when we substitute the expression for the mass of a moving body:

$$p = \frac{m_0 v}{\sqrt{1 - \beta^2}}.$$

---

\* This formula has found experimental confirmation in experiments with  $\mu$ -mesons.

It should be noted that Newton's law remains valid if it is written as  $\mathbf{F} = \frac{d\mathbf{p}}{dt}$ . On the other hand, the formula  $\mathbf{F} = m\mathbf{a}$  will no longer be valid in all cases.

#### Sec. 159. ENERGY

In Sec. 10, we obtained an expression for the energy of a moving body by finding a function that increases as the work expended in accelerating the body. Let us repeat these calculations taking into account the corrections provided by the theory of relativity.

The work of displacing a body by a distance  $d\mathbf{l}$  is

$$\mathbf{F} d\mathbf{l} = \frac{d\mathbf{p}}{dt} d\mathbf{l} = d\mathbf{p} \frac{d\mathbf{l}}{dt} = \mathbf{v} d\mathbf{p},$$

where  $\mathbf{v}$  is the velocity vector. If this work serves to increase the energy  $\mathcal{E}$  of the body, then

$$d\mathcal{E} = \mathbf{v} d\mathbf{p} = v d(mv) = mv dv + v^2 dm.$$

Since  $m = \frac{m_0}{\sqrt{1-\beta^2}}$ , then  $dm = \frac{mv dv}{c^2(1-\beta^2)}$ ; hence,

$$d\mathcal{E} = \left(1 + \frac{v^2}{c^2(1-\beta^2)}\right) mv dv, \quad \text{i.e.,} \quad d\mathcal{E} = \frac{mv dv}{(1-\beta^2)}.$$

Comparing the last expression with the formula for incremental mass, we find:  $d\mathcal{E} = c^2 dm$ , i.e.,

$$\mathcal{E} = mc^2.$$

We have dropped the additive integration constant, for when  $m = 0$ ,  $\mathcal{E}$  must also be equal to zero.

Thus, the work done on a body serves to increase the function  $\mathcal{E} = mc^2$ , which has, therefore, the significance of energy of the body.

The fundamental result of this calculation consists in the following: an increase in the mass of a body is accompanied by an increase in its energy (and, hence, an expenditure of external energy); on the other hand, a decrease in the mass of a body or system is accompanied by a decrease in its energy (and, hence, a transfer of energy to its surroundings). There is a direct and universal relationship between mass increment and energy increment, for  $c^2$  is a constant quantity.

But what is the nature of the energy  $\mathcal{E}$ ? Is it an energy of motion? Evidently not. If the body is at rest,  $\mathcal{E}$  does not equal zero but equals  $m_0c^2$ . Therefore,  $U = m_0c^2$  is the rest energy of the body, i.e., the internal energy of the body, and the difference  $mc^2 - m_0c^2$  is the energy of motion.

The first part of the last sentence should be viewed as an assertion that may be verified experimentally. As for the energy of motion,  $mc^2 - m_0c^2$ , this will be recognised as the familiar expression for the kinetic energy of the following approximation is used:

$$\frac{1}{\sqrt{1-\beta^2}} \approx 1 + \frac{1}{2}\beta^2.$$

To this degree of accuracy,

$$mc^2 - m_0c^2 = m_0c^2 \left( \frac{1}{\sqrt{1-\beta^2}} - 1 \right) \approx m_0c^2 \times \frac{1}{2}\beta^2 = \frac{m_0v^2}{2}.$$

*Example.* The internal energy of a body of mass  $m_0 = 1$  kg is  $U = m_0 c^2 = 9 \times 10^{16}$  J =  $2.16 \times 10^{13}$  kcal. This is the equivalent of the quantity of energy that would be released in the form of heat in the combustion of 3 million tons of coal. Even in thermonuclear reactions, only several per cent of these tremendous reserves of internal energy are released at present.

#### Sec. 160. MASS DEFECT

As was indicated in the preceding article, the expression relating rest mass to the internal energy of a body, i.e.,  $U = m_0 c^2$ , may be verified experimentally.

The internal energy of a body consists of the rest energy of the component parts, their kinetic energy and their potential energy of interaction. A change in any of these component energies affects the value of  $U$  and, hence, the rest mass as well. Thus, the rest mass increases if the temperature of the body rises, i.e., if the internal motion of the system increases. The rest mass also increases if repelling components of the system approach one another or if attracting components move apart.

It is clear from the above that the rest mass of a system of interacting particles does not possess the property of additivity, i.e., it is not subject to the law of conservation. If a body of rest mass  $M_0$  consists of  $N$  particles, each of mass  $m_0$ , then  $M_0 \neq Nm_0$ . The difference

$$M_0 - Nm_0 = \Delta M$$

is called *the mass defect* of the body (or system of particles). The quantity

$$c^2 \Delta M$$

is called *the binding energy*.

If a system breaks up into a number of components, binding energy is released and may be measured. Moreover, the rest mass may also be directly measured. Thus, the  $U = m_0 c^2$  law may be verified experimentally.

Numerical examples show that any change in internal energy related to a change in the velocity of motion and the interaction force between molecules and atoms cannot lead to a measurable change in mass. Experimental verification of this theory is possible in nuclear physics (see p. 426).

*Examples.* 1. The mass of 1 kg of molybdenum increases by  $\Delta M = 0.000000003$  g when it is heated to 1,000 K.

2. If a steel rod of 128 cm length and 1 cm<sup>2</sup> cross-section (mass of the rod = 1 kg) is stretched by a force of 8 tons, the potential energy thereby stored in it increases its mass by  $2 \times 10^{-12}$  gm.

#### Sec. 161. THE PRINCIPLE OF EQUIVALENCE AND THE GENERAL THEORY OF RELATIVITY

Let us consider a noninertial system of coordinates moving with an acceleration  $a_0$ . Assume that we wish to describe physical phenomena in this system. Then laws of mechanics in this system will appear different than in an inertial system, for  $F = ma$  is valid only for the latter. A stationary body will have an acceleration  $-a_0$  relative to this system.

If we maintain the terminology used for an inertial system and assume that acceleration is produced by forces, then the "force" field  $-ma_0$  acting on all bodies in an accelerated system may be called an acceleration field and an analogy may be drawn between this field and a gravitational field.

In exactly the same manner, we may introduce additional "force" fields in considering phenomena in a rotating system of coordinates and, of course, in the general case. The fictitious force fields that we have introduced for the description

of motion from the viewpoint of a noninertial system of coordinates may be called fields of inertial forces. The force  $-ma_0$  is an inertial force.

The motion of a point having an acceleration  $a$  relative to such a noninertial system will obey the equation

$$ma = F + \text{inertial forces.}$$

Expressions for inertial forces may be found in textbooks on theoretical physics.

It is important to direct our attention to the theoretical side of this question. In noninertial systems, fictitious force fields appear. To each such acceleration field there corresponds a fictitious distribution of attracting mass. Hence, any field created by accelerated motion may be interpreted, generally speaking, as a gravitational field. In this sense, we sometimes speak of the equivalence of gravitation and acceleration.

Let us consider several simple examples. Assume that we are in an elevator falling with an acceleration  $a$ . Let us drop a ball and examine the nature of its fall. As soon as the ball is dropped it begins, from the viewpoint of an inertial observer, to fall freely with acceleration  $g$ . Since the elevator is falling with acceleration  $a$ , the acceleration relative to the elevator floor is  $g - a$ . An observer in the elevator can describe the motion of the falling body by means of the acceleration  $g'' = g - a$ . In other words, the observer in the elevator need not speak of the accelerated motion of the elevator since he has "changed" the acceleration of the gravitational field in his system.

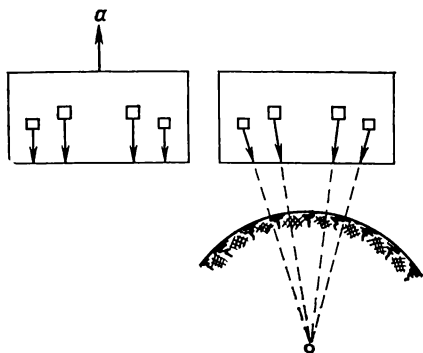


Fig. 185

Now, let us compare two elevators. One is suspended over the Earth and the other moves in interplanetary space with an acceleration  $a$  relative to the stars. All bodies in the elevator suspended over the Earth are able to fall freely with acceleration  $g$ . But bodies inside the interplanetary elevator have a similar capability. They "fall" with an acceleration  $-a$  to the "bottom" of the elevator. The role of bottom is played by the wall opposite to the acceleration direction.

Thus, the action of a gravitational field and the manifestation of accelerated motion are indistinguishable.

The behaviour of a body in an accelerated system of coordinates is the same as the behaviour of a body in the presence of an equivalent gravitational field. However, this equivalence is complete only if we limit ourselves to observations over small portions of space. Thus, imagine an "elevator" having linear floor dimensions of several thousand kilometres. If such an elevator is suspended over the Earth, the phenomena occurring in it will differ from those occurring in an elevator moving with an acceleration  $g$  relative to fixed stars. This is clear from Fig. 185. In one case bodies fall obliquely to the bottom of the elevator, while in the other case perpendicularly.

Thus, the principle of equivalence is valid for such volumes of space in which the field may be considered uniform.

The above qualitative considerations lie at the basis of the general theory of relativity. This theory was also developed by Einstein. In it, he sought formulations

for the laws of nature, independent of the choice of coordinate system. Until now, we have assumed that this was possible only for inertial systems of coordinates. The principle of equivalence shows that the absoluteness of acceleration may be destroyed by a gravitational field. An accelerated system of coordinates may be viewed as an inertial system if we introduce an equivalent gravitational field. To be sure, as we have just seen, such equivalence is limited in time and space. However, Einstein showed that this restriction may be removed if, corresponding to the introduction of the gravitational field, a change in the geometry of the system is introduced.

# The Quantum Nature of a Field

## Sec. 162. PHOTONS

On a number of occasions we have indicated that radiation and absorption of electrical energy occur in packets, or quanta. The magnitude of a quantum depends only on its radiation frequency and is equal to  $h\nu$ , where  $h$  is a universal constant equal to  $6.62 \times 10^{-27}$  erg s. It should be noted that the quantum nature of radiation and absorption has been established already for the entire electromagnetic spectrum, i.e., from hard  $\gamma$ -rays to long radio waves.

The phenomena of radiation and absorption characterise, in the first place, the microsystem interacting with the electromagnetic field of a wave. The quantum nature of these phenomena (which we shall discuss in detail in Part III) shows that a microsystem has distinct energy levels and that the values of these energy levels cannot be arbitrary. These facts by themselves would not have led to the conclusion that this quantum nature is characteristic of an electromagnetic field as well as of matter if an electromagnetic wave in its interaction with matter did not behave, in a number of cases, as a particle. The corpuscular properties of electromagnetic radiation are manifested when losses and transformations of electromagnetic energy occur. The shorter the wavelength, the more distinct the effects. These properties, on the other hand, are not manifested during propagation, scattering and diffraction of electromagnetic waves if these processes are not accompanied by energy losses.

A corpuscle of an electromagnetic field is called a *photon*. It is characterised, in the first place, by the magnitude of its energy:

$$\mathcal{E} = h\nu.$$

Using the law of equivalence of mass and energy, we are entitled to ascribe to a photon the mass

$$m = \frac{\mathcal{E}}{c^2} = \frac{h\nu}{c^2}.$$

Since an electromagnetic field is propagated with a velocity  $c$ , it must be concluded from the formula  $m = \frac{m_0}{\sqrt{1-\beta^2}}$  that the rest mass of a photon is equal to zero.

Assuming the concept of momentum applicable to a photon, we obtain

$$p = mc = \frac{h\nu}{c}.$$

It should be recalled that the Lebedev experiments (see p. 242) directly demonstrate the validity for light of the formula  $p = \frac{W}{c}$ , the relationship between the momentum density and energy density of an electromagnetic wave. The formula for photon momentum is in complete agreement with this result.

As may be seen from what follows numerous experiments convincingly show that photons exist. On the other hand, a mass of experimental evidence prevents us from abandoning our view of an electromagnetic field as a continuum. The sharpest contradictions arise in considering interference phenomena. These phenomena are elegantly explained by the wave nature of the field, but are completely inexplicable from the corpuscular viewpoint.

Values of  $\mathcal{E}$ ,  $m$  and  $p$  for Photons of Various Types  
of Electromagnetic Radiation

	$\lambda$	$\mathcal{E}$	$m$ (g)	$p$ (g/cm/sec)
Radio waves . . . . .	2,000 m	$10^{-21}$ erg = $0.62 \times 10^{-9}$ eV	$1.1 \times 10^{-42}$	$3.3 \times 10^{-32}$
Visible light . . . . .	6,000 Å	$3.3 \times 10^{-12}$ erg = 2 eV	$3.6 \times 10^{-33}$	$1.1 \times 10^{-22}$
X-rays . . . . .	1 Å	$19.8 \times 10^{-9}$ erg = 12,400 eV	$2.2 \times 10^{-29}$	$6.6 \times 10^{-19}$

Thus, consider a simple interference arrangement—two close apertures through which light may be transmitted. The following experiment is easily performed. First, let us photograph the transmitted light when both apertures are open. We obtain the pattern discussed earlier, namely, alternate bright and dark fringes. Now, let us close each of the apertures in succession and take the photograph on one plate. The result, of course (from the viewpoint of wave theory), will be different, i.e., there will be no interference.

Let us now consider how this experiment may be interpreted in corpuscular terms. It is conceivable (by stretching one's imagination) that photons fall unequally on different parts of the photographic plate owing to rebound from the edge of the aperture or collision with one another. But the patterns obtained differ depending on whether the light passes through both apertures simultaneously or consecutively. Photons passing through one aperture "know" whether the other aperture is open or closed.

This experiment and many others show that it is quite impossible to reduce electromagnetic phenomena to only a field pattern or to only a system of photons. Each concept is exceedingly fruitful in the case of one group of phenomena, but fails in the case of the other.

During the last few decades, physicists have energetically sought ways of reconciling these two contradictory views of electromagnetic radiation. A field is a reality characterised by continuous values of field intensity in space and time; a corpuscle is a reality occupying a certain limited region of space at a given instant. These contradictory qualities are combined in electromagnetic radiation. In Chapter 27, we shall see that these contradictory properties are combined not only in the case of electromagnetic radiation, but in the case of matter as well. However, physics has made considerably more progress in understanding matter than in understanding an electromagnetic field. The dual nature of particles of matter is described by the Schrödinger equation (see p. 369); interactions between corpuscles and waves for such particles are understood quite well.

Unfortunately, the situation is much worse as regards electromagnetic field (radiation) theory, commonly referred to as quantum electrodynamics (for a detailed discussion, see p. 451). Such a complete theory does not exist. In view of the fundamental contradictions existing in quantum electrodynamics, its partial successes, expressed in the establishment of new relationships between field and particles, cannot be generalised. Hence, the interrelation between photons and electromagnetic field remains unclear.

The rules of "translation" from corpuscular terminology to wave terminology, and vice versa, are based on the following: an electromagnetic wave of length  $\lambda$  and intensity  $I$  may appear as a stream of photons of frequency  $\nu = \frac{c}{\lambda}$  and intensity  $I = N h \nu$ , where  $N$  is the number of photons passing per unit time through

unit area. The direction of motion of the wave front is the direction of motion of the photon.

We shall not discuss in corpuscular terms the very complex problem of the polarised state of light. To do this, it is necessary to assume that a photon has a selected direction, or spin (see p. 386 regarding electron spin).

### Sec. 163. PHOTOELECTRIC EFFECT

The escape of electrons under the action of electromagnetic waves constitutes important confirmation of the indispensability of the corpuscular viewpoint. This phenomenon will be considered here from this viewpoint, and again in Sec. 287 when we discuss the action of light on metals and semiconductors.

Since the escape energy of an electron from a metal (see p. 550) is not less than 2.2 eV, the photoelectric effect becomes possible when  $h\nu > 3.5 \times 10^{-12}$  erg, i.e., for frequencies of the order of  $0.5 \times 10^{15}$  Hz ( $\lambda = 6,000 \text{ \AA}$ ).

Einstein proposed that the photoelectric effect be viewed as an effect of collision between a photon and an electron. In this process, the photon gives up all of its energy and ceases to exist. If  $A$  represents the work function of electron, i.e., the work required to overcome the binding force between the electron and the substance, the law of conservation of energy has the form

$$h\nu = A + \frac{mv^2}{2},$$

where  $\frac{mv^2}{2}$  is the kinetic energy of the photoelectron, the electron dislodged from the substance.

The first means of checking the validity of the photon hypothesis consists in verifying the linear dependence between photoelectron kinetic energy and frequency of incident radiation.

The photoelectron energy is determined by the bias potential method. If the surface of the substance from which the electrons are dislodged constitutes a condenser plate, current flows through the circuit in which this condenser is connected. Current ceases to flow when an appropriate bias voltage is applied to the condenser. This condition is given by

$$eU_b = \frac{mv^2}{2}.$$

It should be realised that the greater the depth from which the electrons are dislodged, the smaller the velocities. Therefore, current ceases to flow when the electrons closest to

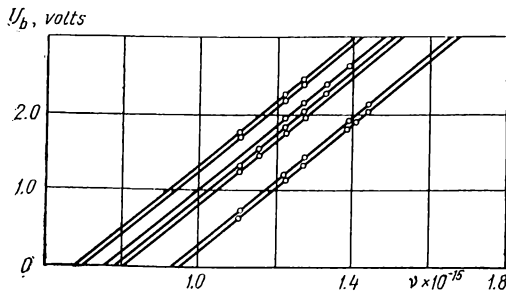


Fig. 186

the surface are prevented from escaping. By experimentally determining  $eU_b$  for various frequencies  $\nu$  of electromagnetic radiation, curves of  $U_b$  vs.  $\nu$  may be plotted. The ideal straight lines obtained are shown in Fig. 186.

The slope  $\frac{h}{e}$  of the straight line  $eU_b = h\nu - A$  may be calculated from other data, providing another independent means of checking the validity of the theory.

Nevertheless, the above experiment cannot be considered the direct proof of the photon hypothesis. The possible objection is that the photoelectron may gradually



accumulate the energy transmitted to it by an electromagnetic wave. This objection was answered by the classical experiment of A.F. Yoffe and N.I. Dobronravov. Earlier, Yoffe had investigated the photoeffect by means of a particle of dust suspended between the plates of a condenser. Owing to inevitable air friction, the particle of dust carries a charge and, hence, its weight may be counterbalanced by an electric field. For equilibrium  $qE = mg$ , where  $m$  is the mass and  $q$  is the charge of the particle of dust. In the photoeffect process, the particle of dust loses an electron and, hence, depending on the sign of  $q$ , changes its charge by  $q + e$  or  $q - e$ . The particle of dust is then no longer in equilibrium and begins to move towards one of the condenser plates. To counterbalance the dust particle, it is necessary to change the field. The equilibrium condition is now

$$(q \pm e) E_1 = mg.$$

In this manner, Yoffe determined the charge of an electron.

Now, let us describe the Yoffe and Dobronravov experiment. Here, too, the behaviour of a dust particle suspended between the two plates of a condenser was observed, but now the goal was different. The anode of an X-ray tube served as one of the condenser plates. A voltage of 12,000 V was applied to the tube and the X-rays were created by an exceedingly weak electron flux of about 1,000 electrons per second.

As is well known, X-rays are created when an electron strikes an anode. But what is radiated by the anode? Is it a continuous electromagnetic field or 1,000 photons per second? The dust particle between the condenser plates enables us to obtain the answer. The X-rays dislodge electrons from the dust particle. But how do they do this?

The Yoffe and Dobronravov experiment showed that, on the average, one electron was dislodged from the dust particle every 30 minutes. If the X-rays were propagated in the form of a continuous field, then at each instant the dust particle would have obtained a very minute amount of energy, insufficient of course to dislodge an electron. This energy would have been evenly distributed among all the electrons of the dust particle. From the wave viewpoint, a quite inconceivable conclusion would have to be drawn from the Yoffe and Dobronravov observations, namely, that once every 30 minutes all the electrons transfer energy to one electron, which then escapes from the dust particle.

The photon hypothesis not only explains the phenomenon qualitatively but quantitatively as well. The dust particle in the above experiment consisted of a bismuth spherule having a radius of  $3 \times 10^{-5}$  cm. It was located at a distance of 0.02 cm from the anode, from which X-rays emerged in all directions. The probability of a photon striking the dust particle is  $\frac{\pi (3 \times 10^{-5})^2}{4\pi (0.02)^2} = \frac{1}{1,800,000}$ . Since in 1 second 1,000 photons are dislodged, on the average 1 photon will strike the dust particle every 1,800 sec (30 minutes), which agrees with the experimental result.

#### Sec. 164. FLUCTUATIONS IN LUMINOUS FLUX

The experiments of S.I. Vavilov devoted to the study of fluctuations in luminous flux of low intensity provide important experimental corroboration of the photon theory.

It turns out that the eye's threshold of sensitivity to light is exceedingly low. The human eye is capable of perceiving approximately 100 photons per second falling on the cornea. If the luminous flux fluctuates about this value, light will

not be perceived by the eye when the number of photons drops below the threshold value.

In the Vavilov experiments, the investigator observed a beam of light that was discharged every second for a time interval of 0.1 sec. When the value of the luminous flux exceeded the threshold of sensitivity, the eye perceived every flash of light. When the light intensity was decreased, some of the flashes were no longer perceived by the observer. The lower the light intensity, the greater the number of flashes that were not perceived. Thus, fluctuations in the number of photons in the luminous flux were directly observed. It is difficult to provide more direct evidence of the corpuscular nature of light.

Other experiments performed by Vavilov clearly show that such typically wave phenomena as interference cannot be explained by the photon hypothesis. Using a Fresnel biprism, Vavilov divided a beam of light into two coherent components. These components yielded an interference pattern. At the same time, fluctuations in both beams of light were completely independent. This circumstance again shows that it is quite impossible to explain interference as some statistical distribution of photons.

Wave properties are inherent in every photon rather than in a stream of photons. Thus, a photon can in no way be viewed as an "ordinary" particle.

At this point, we must digress somewhat. In creating a model of the invisible world, we endow elementary particles with properties borrowed from the world of things (materials) around us, or, as they say in physics, from the macroworld. Thus, for example, atoms are conceived as spherules. Needless to say, an atom spherule only partially reflects the properties of a material spherule. Everyone knows, for example, that such properties inherent in a material spherule as colour, roughness and odour cannot be transferred to an atom spherule. The more we penetrate into the microworld, the more difficult it is to endow elementary particles with material properties.

Components of an atom or atomic nucleus and particles of light resemble a material spherule even less than an atom does. In the case of a photon, we saw that it is possible to combine in a microparticle conflicting properties of the macroworld. Of course, in the macroworld, a particle is a particle and a wave is a wave. A particle occupies a limited region of space and travels along a definite path. A wave is distributed continuously in space and the energy is transferred to one or another region from all points in space. For materials, these two views are irreconcilable. But we have no right to impose the behaviour of materials on particles of the microworld.

Cognition of the microworld does not consist in the creation of a model resembling the pictures familiar to the human eye. The infinite process of cognition consists in the investigation of the regularities of phenomena, the determination of objectively existing causal relationships between phenomena. In this manner, a complex picture of the microworld, whose essence cannot be transmitted by any ingenious model borrowed from the macroworld, is obtained.

#### Sec. 165. KIRCHHOFF'S LAW

It has been experimentally established that two bodies having different temperatures tend to equalise their temperatures even when the bodies are in a vacuum. The energy exchange occurs by means of electromagnetic waves radiated by the atoms of these bodies.

As was indicated above, a specific system of energy levels is associated with every atom. When an atom absorbs energy, its energy level rises; when it radiates,

its energy level decreases. During every radiation process, an atom releases into space an electromagnetic energy  $h\nu_{mn} = \mathcal{E}_m - \mathcal{E}_n$ , where  $\mathcal{E}_m$  is the energy level before radiation and  $\mathcal{E}_n$  the level after radiation. The radiated wave has a frequency  $\nu_{mn}$ . This wave arrives at the other body and is absorbed by it. In this case, the energy level of the atom absorbing energy is raised from  $\mathcal{E}_n$  to  $\mathcal{E}_m$ .

The same thing may be expressed in terms of photons. Thus, it may be stated that during every radiation process a photon of electromagnetic energy  $h\nu$  is released; during absorption the photon is captured by an atom and its energy serves to raise the energy level of the atom.

All the atoms of the bodies participate in the energy exchange—sometimes a photon is absorbed, sometimes a photon is radiated. Depending on the random circumstances, the most varied energy transitions may occur and, in principle, electromagnetic waves of any wavelength may participate in the energy exchange.

Let us assume that the bodies participating in the heat exchange form a closed system, i.e., the system of bodies under observation is surrounded by an envelope that prevents radiation from passing through. Then, after a certain interval of time, these bodies reach a state of equilibrium and assume the same temperature. This does not mean that electromagnetic radiation ceases. As before, a transition will sometimes occur to a higher energy state of an atom and sometimes to a lower. But if the equilibrium state has been reached, then for each body, at each instant of time, equal quantities of energy will arrive and leave. This is true for radiation of any wavelength. In general, the radiation arriving at a body is only partially absorbed, raising the energy levels of its atoms from the lowest energy level to the highest one. The other part of the incident radiation is scattered, i.e., reflected, by the body.

Atoms do not maintain their high energy levels long: in returning to their original state, they give up the absorbed energy in the form of radiation. If the energy incident on a unit area in 1 sec is designated by  $\rho$ , the absorbed energy may be expressed as  $A\rho$ . The dimensionless coefficient  $A$ , indicating the fraction of energy that is absorbed, is known as *the absorptivity* of the body. Evidently, if

$$A\rho = \mathcal{E},$$

where  $\mathcal{E}$  is the energy radiated from 1 cm<sup>2</sup> of surface in 1 sec, the body is in equilibrium with its surroundings and its temperature does not change.

But what is the condition for thermal equilibrium of many bodies, which may have, of course, different absorptivities and different radiations? On the basis of thermodynamical considerations, Kirchhoff showed that equilibrium is possible only if the intensity of the electromagnetic waves incident on a body is the same for all portions of the bodies in equilibrium with one another. Thus,

$$\frac{\mathcal{E}_1}{A_1} = \frac{\mathcal{E}_2}{A_2} = \frac{\mathcal{E}_3}{A_3} = \dots = \rho.$$

This relationship is known as *Kirchhoff's law* and is valid for any wavelength and any temperature. It states that the ratio of the emissive power of a body to its absorptivity is a constant for any wavelength and temperature.

This means that a body that is a good absorber of certain rays is also a good radiator of these rays, and vice versa. Why does the temperature of water in a bottle coated with silver rise slowly, and the temperature of water in a dark flask rise rapidly, under the action of solar rays? In the first case there is little absorption of solar energy, while in the second there is considerable absorption. Now, let us assume both vessels are filled with hot water and placed in a refrigerator.

The water in the dark flask cools much more rapidly since the better absorber is also a better radiator.

A striking experiment may be performed with coloured ceramic. If the colour of a body is, for example, green, it will *not* absorb green light. Thus, if we heat a green crock, it is seen that it begins to assume a colour complementary to green.

It should not disturb us that we have applied a law established for equilibrium to phenomena involving bodies clearly not in equilibrium (the body is at a higher temperature than its surroundings). The situation here is exactly the same as in the case of other thermodynamic problems (cf. p. 123): the laws of thermodynamics are applicable if every instantaneous state may be viewed as an equilibrium state. In thermal radiation phenomena, this condition is always satisfied.

#### Sec. 166. BLACK-BODY RADIATION

Kirchhoff's law has an interesting consequence. Bodies exchanging heat by means of radiation receive (for given values of  $\nu$  and  $T$ ) the same electromagnetic wave intensity from their neighbours, independent of the material and properties of the bodies. For every wavelength (or, what amounts to the same, every frequency) and for every temperature, experiments yield a universal value for  $\rho$ .

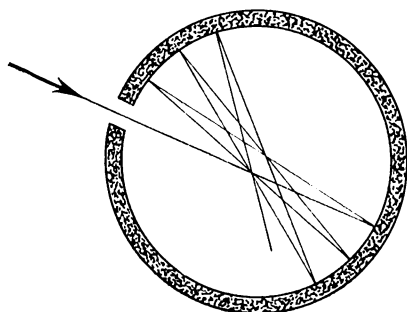


Fig. 187

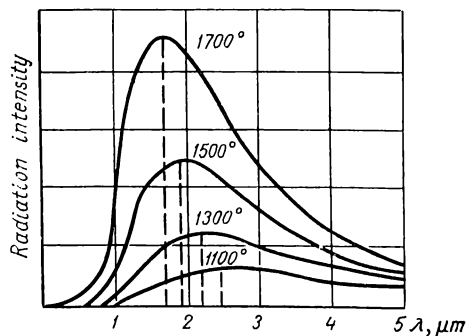


Fig. 188

Thus, there exists a universal function  $\rho(\nu, T)$ , i.e., a function of the radiation frequency and temperature, characterising the process of thermal exchange by radiation.

The meaning of the function  $\rho(\nu, T)$  is easily explained. Consider a body absorbing 100 per cent of the energy incident on it for all wavelengths. For such a *perfectly black* body,  $A = 1$  and

$$\mathcal{E} = \rho(\nu, T).$$

The function  $\rho(\nu, T)$  is the *emissive power* of a perfectly black body. But what kind of body absorbs light of all wavelengths? Of course substances such as lampblack (soot) are almost perfectly black. However, all such substances fall short by several per cent of the condition  $A = 1$ . A more ingenious solution exists. Imagine a box having a small aperture. If the aperture's dimensions are made sufficiently small, it may be made perfectly black. This property of apertures is well known from everyday observations. A deep hole, an open window nonilluminated from within the room and a well are examples of perfectly black "bodies". It is clear what happens in these cases: a beam entering a cavity through an aperture is able to emerge only after repeated reflections (see Fig. 187). But with each reflection, part of the

energy is lost. Therefore, in the case of a small aperture and a large cavity, the beam is unable to emerge, i.e., it is completely absorbed.

To measure the emissive power  $\rho(\nu, T)$  of a perfectly black body, a long tube made of refractory material is placed in an oven and heated. Through an aperture in the tube, the nature of the radiation is studied by means of a spectrograph. The results of such experiments are shown in Fig. 188. The radiation intensity is plotted as a function of the wavelength for several temperatures. It is seen that the radiation is concentrated in the relatively narrow spectral interval of 1 to 5  $\mu\text{m}$ . Only at high temperatures do such curves take in portions of the visible spectrum and begin to advance in the direction of short waves. Waves whose wavelengths are several microns long are called infrared waves. Since for ordinary temperatures they are the main carriers of energy, we call them heat waves.

The higher the temperature, the more distinct the maximum of a thermal radiation curve. With increasing temperature, the wavelength  $\lambda_m$  corresponding to the maximum of the spectrum is displaced in the direction of shorter wavelengths. This displacement obeys the Wien law, which is easily established experimentally:

$$\lambda_m = \frac{2,886}{T}.$$

In this formula, the wavelength is expressed in microns and  $T$  in degrees Kelvin. The displacement of radiation in the direction of shorter wavelengths may be detected when a metal is heated. As the temperature increases, the colour of the heat changes from red to yellow.

We call the reader's attention to another feature of the radiation curves, namely, it will be noted that all the ordinates increase sharply with increasing  $T$ . If  $\mathcal{E}_\lambda$  is the intensity for a given wavelength, the total intensity of the spectrum is expressed by the integral

$$R = \int_0^\infty \mathcal{E}_\lambda d\lambda.$$

This integral is simply equal to the area under the radiation curve. Exactly how rapidly does  $R$  increase with increasing  $T$ ? Analysis of the curves shows that it increases very rapidly, namely, proportionally to the fourth power of the temperature:

$$R = \sigma T^4 \text{ ergs/cm}^2\text{sec},$$

where  $\sigma = 5.7 \times 10^{-5}$  (CGS units). This is the Stefan-Boltzmann law.

Both laws are important in the determination of the temperature of hot bodies at great distances. In this manner, the temperature of the Sun, stars and the ball of fire of an atomic explosion are determined.

The laws of thermal radiation are basic to the determination of the temperature of smelted metal. The operation of optical pyrometers is based on the selection of the heating for an electric bulb filament in such a manner that the luminosity of this filament becomes the same as the luminosity of the smelted metal. We make use of the following law: if radiations are the same, so are the temperatures. As for the temperature of the heated filament, it is directly proportional to the electric current flowing through the filament. Hence, it is not difficult to calibrate an optical pyrometer. Since actual bodies are not perfectly black, it is necessary to introduce in each case in the Stefan-Boltzmann formula a factor less than unity (the absorptivity of the given body). These factors are determined empirically and are significant in heat engineering where problems of heat exchange by radiation are extremely important. Nevertheless, the above laws are valuable since the

general behaviour of radiation (dependence on temperature and wavelength) is maintained for bodies that are not black as well. The theoretical aspects of black-body radiation are discussed in the following article.

#### Sec. 167. THE THEORY OF THERMAL RADIATION

Let us consider a cavity within which absorption and radiation of electromagnetic waves occur. It is immaterial whether this cavity is in the form of a sphere or a rectangular parallelepiped. The walls of the cavity radiate and absorb equal quantities of energy, i.e., the entire system is in equilibrium. Within the cavity there is an electromagnetic field which is in equilibrium with the walls: at all points the energy density of the field,  $w = \frac{1}{8\pi} (E^2 + H^2)$ , is constant in time.

This electromagnetic field may be viewed in two different ways. From one viewpoint, there are standing electromagnetic waves in the cavity, just as there are standing sound waves in a closed room with sound sources. From the other viewpoint, in view of the quantum nature of the field, it may be stated that the space under consideration is filled with photons, just as a vessel containing gas is filled with molecules.

From the wave viewpoint, the number of frequencies of electromagnetic oscillations occurring in the cavity may be easily determined. The reasoning used for sound waves (see p. 102) is completely applicable here too. The number of characteristic frequencies of electromagnetic oscillations less than  $\nu$  is equal to

$$\frac{4}{3} \pi \frac{\nu^3}{c^3} V,$$

where  $c$  is now the velocity of electromagnetic waves and  $V$  is the volume of the cavity. This formula gives the number of oscillations for the case of linearly polarised waves. In the case of thermal radiation, we are dealing with nonpolarised oscillations, which may be always resolved into components along two axes.

Here, the number of oscillations is twice as large and is equal to  $\frac{8}{3} \pi \frac{c^3}{\nu^3} V$ . Differentiating, we obtain the number of oscillations in the frequency interval from  $\nu$  to  $\nu + d\nu$ :

$$\frac{8\pi\nu^2}{c^3} V d\nu.$$

Now, let us consider the situation from the "other side of the coin". From this viewpoint, the cavity is filled with oscillations of frequency  $\nu$ —in other words, with photons of energy  $\varepsilon = h\nu$ . The expression  $\frac{8\pi\nu^2}{c^3} V d\nu$  may be viewed as the number of photons in the cavity, and  $\frac{8\pi\nu^2}{c^3} d\nu$  as the density of the photon gas. We shall soon be able to answer the following important question: what is the electromagnetic energy density in the cavity? If photons of all energies were created in equal numbers, it would merely be necessary to multiply  $\varepsilon$  by  $\frac{8\pi\nu^2}{c^3} d\nu$  to obtain the energy density for frequencies in the interval  $d\nu$ . However, the particle energies are not distributed uniformly. Therefore, the formula being sought has the form

$$w_\nu d\nu = \frac{8\pi\nu^2}{c^3} \varepsilon W(\varepsilon) d\nu,$$

where  $W(\varepsilon)$  is the probability of a photon of energy  $\varepsilon$  being created.

Thus, the electromagnetic energy density for waves (photons) of frequency  $\nu$  is given by the formula

$$w_{\nu} = \frac{8\pi\nu^2}{c^3} \varepsilon W(\varepsilon).$$

The energy flux through a unit area, i.e., the Poynting vector  $K$ , is  $c$  times  $w_{\nu}$  (see p. 245). But the energy flux  $\rho$  radiated from a unit area of a body in equilibrium with the field is one-fourth of the value of the Poynting vector:  $\rho = \frac{1}{4} K$ . Thus, between  $w$  and  $\rho$ , the following relationship exists:  $\rho = \frac{c}{4} w$ . What is the origin of the coefficient  $\frac{1}{4}$ ? Since, on the whole, this is not a very important matter, the following simplified explanation should suffice.

Every unit area radiates an energy flux  $\rho$  in all directions within the limits of a hemisphere, i.e., a solid angle  $2\pi$ . Thus, the average radiation within a unit solid angle is equal to  $\frac{\rho}{2\pi}$ . From geometry considerations, it is clear that the radiation is equal to zero in the plane of a unit area and a maximum along its normal. If the decrease in radiation intensity were uniform, to obtain the average value  $\frac{\rho}{2\pi}$  it would be necessary for the radiation along the normal to be equal to  $\frac{\rho}{\pi}$ .

Now, consider a sphere filled with radiation. At the centre of the sphere there is a unit area through which there is an energy flux  $K$ . On the other hand, however, the radiation falling on this area from all parts of the sphere is equal to  $\frac{\rho}{\pi} \times 4\pi$ . Hence,  $\rho = \frac{1}{4} K$ .

Thus, using the formula for the volume density of electromagnetic radiation, we obtain an expression for the emissive power of a black body by multiplying  $w_{\nu}$  by  $\frac{c}{4}$ :

$$\rho_{\nu} = \frac{2\pi\nu^2}{c^2} \varepsilon W(\varepsilon).$$

Further investigation of this function involves evaluation of  $W(\varepsilon)$ , the energy distribution probability. Historically, the first formula for  $\rho_{\nu}$  was proposed in 1911 by Rayleigh and Jeans independently of each other. It has the following form:

$$\rho_{\nu} = \frac{2\pi kT}{c^2} \nu^2.$$

This formula was obtained by assuming iniform distribution of energy per degree of freedom, i.e.,  $W$  independent of  $\varepsilon$ . It is valid for long wavelengths and high temperatures.

Another possibility for  $W(\varepsilon)$  is to use the Boltzmann law, which was so successful in the case of molecular gases. Then,  $W(\varepsilon) = e^{-\frac{\varepsilon}{kT}}$ . However, as may be seen from Fig. 189, both the Wien emissive power formula,

$$\rho_{\nu} = \frac{2\pi\nu^2}{c^2} h\nu e^{-\frac{h\nu}{kT}},$$

and the Rayleigh-Jeans formula do not agree with experimental results.

Where is the fallacy in reasoning in these cases? It must be sought in the inapplicability of the statistical reasoning lying at the basis of Boltzmann's law to an aggregate of photons. As we have already emphasised, photons give us a one-sided

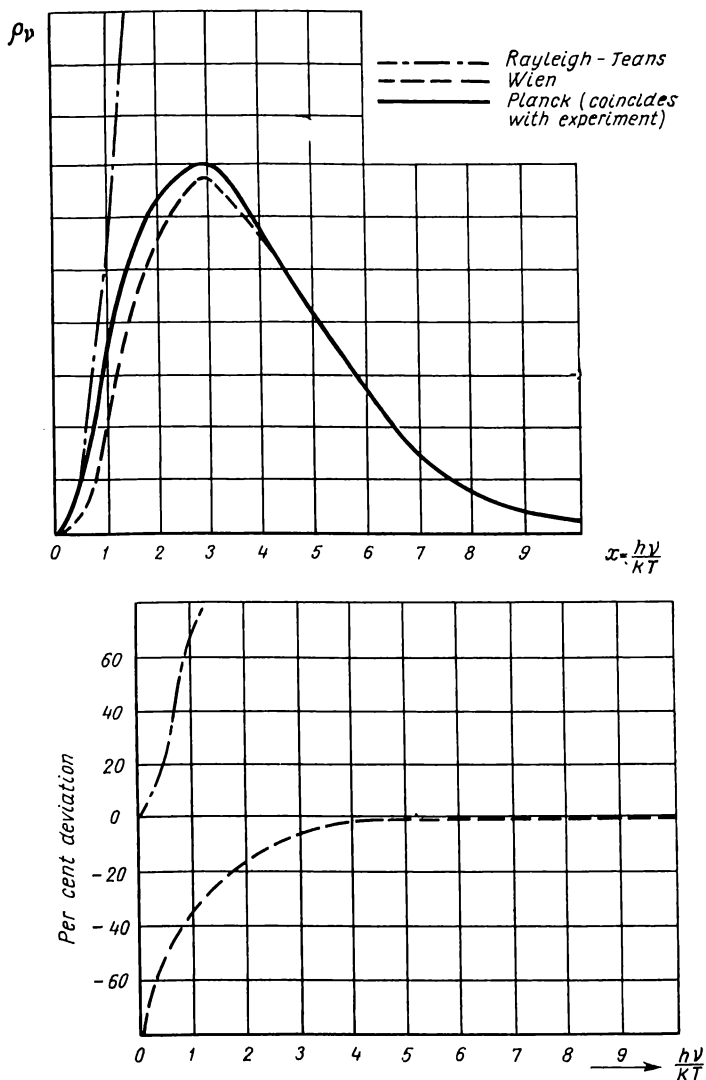


Fig. 189

picture of an electromagnetic field. The reality of the field cannot be completely represented by a collection of particles. It is, therefore, natural that photons should have their "own statistics" (Bose-Einstein statistics).

To obtain the new function of particle energy distribution replacing the Boltzmann law, all we need do is take proper account of the fact that the wave nature of the field makes the concept of a difference between identical particles meaningless. Bose-Einstein statistics is founded on this new basis (see p. 542). It leads to the following law of photon energy distribution:

$$W(\epsilon) = \frac{1}{e^{\frac{h\nu}{kT}} - 1}.$$



Therefore, the formula for the emissive power of a perfectly black body has the following form:

$$\rho_\nu = \frac{2\pi\nu^2}{c^2} \frac{h\nu}{e^{h\nu/kT} - 1}.$$

This formula was first obtained by Planck and is named after him. The excellent agreement between this theoretical formula and experimental results, and the nature of the deviations of the Wien and Rayleigh-Jeans formulas, are illustrated in Fig. 189.

The Wien and Stefan-Boltzmann laws considered above follow from Planck's formula. To prove the first of these laws, it is necessary to solve the problem for an extremum, i.e., find the root of the equation

$$\frac{\partial \rho_\nu}{\partial \nu} = 0.$$

To prove the second law, it is necessary to find

$$\int \rho_\nu d\nu.$$

We leave these calculations to the reader.

#### Sec. 168. STIMULATED EMISSION OF RADIATION

Let us return to the cavity within which there is an electromagnetic field which is in equilibrium with the walls of the cavity. But now we shall view this system from the microscopic viewpoint. Here, photons are emitted by excited atoms. Let us concentrate on  $h\nu$ -photons. The gas of these photons is in equilibrium with the atoms which emit and absorb light with the frequency  $\nu$ , i.e. with the atoms possessing the energies  $E_2$  and  $E_1$ , where  $E_2 - E_1 = h\nu$ .

When the equilibrium is settled, the numbers of atoms  $N_1$  found in level  $E_1$  and atoms  $N_2$  found in level  $E_2$  will remain unchanged. Since the distribution of atoms according to energies obeys the Boltzmann law, we have

$$\frac{N_1}{N_2} = e^{(E_2 - E_1)/kT} = e^{h\nu/kT}.$$

Equilibrium of the system is, of course, of the dynamic character, i.e. the atoms jump from a lower level into an upper level, as also in the reverse direction, photons being now absorbed, now emitted. Since equilibrium takes place, the numbers of transitions in both directions per unit time are equal to each other.

Two processes seem to be obvious here. The first of them is the absorption of a photon which occurs when the latter meets an atom in the lower level  $E_1$ ; as a result the atom gets "excited", i.e. it jumps into level  $E_2$ . The number of such events can be written in the form

$$BN_1W(\epsilon).$$

Here  $B$  is a proportionality factor, and our formula states that the number of transitions of the atoms from a lower level to an upper one (i.e., the number of photon absorptions) is proportional to the number of atoms of energy  $E_1$  and to the number of photons of energy  $\epsilon$ .

The second process, whose existence is obvious, consists in spontaneous transitions of the atoms from an upper level to a lower level. Since the position of an atom in level  $E_2$  is unstable, they will gradually jump into the lower level. The number of such transitions per unit time must be proportional to the number of

excited atoms in the system:

$$AN_2,$$

where  $A$  is another proportionality factor.

If we equate  $AN_2$  to  $BN_1W(\epsilon)$ , then for  $W(\epsilon)$  we shall obtain the Boltzmann distribution law, i.e. the same statistics which is obeyed by the atoms. But this assumption, as we saw in the preceding section, leads to a sharp contradiction with the experiment. Hence, in addition to these two processes, there is one more which takes part in creating the above mentioned balance.

After Planck had published his formula, Einstein pointed out at once that everything would be explained if we assumed that falling onto an excited atom, a photon of energy  $h\nu$  stimulates its emission with the same frequency. Also, the probability of this process must be the same as the probability of absorption, i.e. the number of the acts of stimulated emission per unit time is equal to

$$BN_2W(\epsilon).$$

Equating now the numbers of transitions in opposite directions, we get

$$AN_2 + BN_2W(\epsilon) = BN_1W(\epsilon),$$

or

$$W(\epsilon) = \frac{A/B}{e^{h\nu/kT} - 1}.$$

The limiting values of  $W(\epsilon)$ , namely  $W(\epsilon) = e^{hT/h\nu}$  for small  $h\nu/kT$  (Rayleigh-Jeans) and  $W(\epsilon) = e^{-h\nu/kT}$  for large  $h\nu/kT$  are known to us. This causes us to set  $A = B$ . In such a way we demonstrate the way to the formula for photon statistics

$$W(\epsilon) = \frac{1}{e^{h\nu/kT} - 1}.$$

We see that the spectrum of a black body is explained only by introducing the concept of stimulated emission of radiation.

We can show that stimulated emission of radiation must essentially differ from spontaneous emission. Spontaneously emitted photons have different directions and random phases, while those which came into existence due to the meeting of the  $h\nu$ -photon with an excited atom have the same phase and the same direction as the primary photon. Thanks to these features of stimulated emission of radiation, we can obtain fantastic powers of light fluxes in devices called *lasers* (for Light Amplification by Stimulated Emission of Radiation), or *quantum-mechanical oscillators* (or *amplifiers*)

## Sec. 169. LUMINESCENCE

We speak of luminescence when molecules can be brought to an excited state without increasing their average kinetic energy, i.e. without heating.

Luminescence does not obey Kirchhoff's law. Luminescence intensity, by definition, does not exceed the intensity of radiation of the same wavelength emitted by an absolutely black body.

We distinguish between two types of luminescence: fluorescence and phosphorescence. A phenomenon is called fluorescence (from the Latin *fluor* for flow) if it consists in a spontaneous transition of a molecule from the excited state  $F$  into the lower level  $N$  (Fig. 189a). The duration of fluorescence is usually less than  $10^{-7}$  sec and in any case less than one second.

If an excited molecule or atom jumps from an excited level into a metastable level, then phosphorescence may occur (from the Greek *phos* for light and *phoros* for carrying). Metastable is the name given to such a level transitions from which to a lower level are hardly probable. Emission now can be caused only by a molecule returning from level  $M$  to the previous excited level  $F$ . At high temperatures the molecule returns rapidly, at lower temperatures it does slowly. Thus, phosphorescence, in contrast to fluorescence, depends on temperature.

Luminescence can be caused by various factors: by a chemical reaction, by friction, and so on. The principal types of luminescence are photoluminescence and electroluminescence occurring by virtue of light absorption and impacts by charged particles.

Photoluminescence cannot have a higher frequency (i.e. a greater energy quantum) than the exciting light. Red fluorescence is excited by orange light, yellow by green, green by blue, and so on.

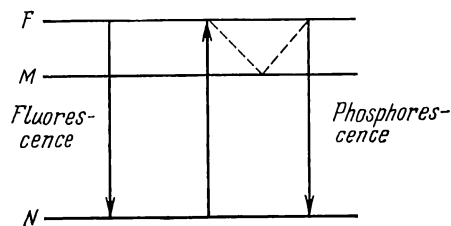


Fig. 189a

## PART THREE

# Structure and Properties of Matter

## CHAPTER 26

### Streams of Charged Particles

The simplest form of matter is an aggregate of charged elementary particles—electrons and ions. Systems of charged particles are encountered in the form of beams of particles in which all the particles have a common velocity and move in a single direction and in the form of a gas in which the particles move randomly. Intermediate states are, of course, also possible. In this chapter, we shall consider the basic physical phenomena of such systems and describe the equipment in which beams and gases of charged particles are used. Problems of electron emission, directly related to the solid state physics, will be discussed in Chapter 37.

#### Sec. 170. MOTION OF CHARGED PARTICLES IN ELECTRIC AND MAGNETIC FIELDS

A force  $f = eE + \frac{e}{c} [vB]$  is exerted on a charged particle in an electromagnetic field (see p. 201). If the fields  $E$  and  $B$  are given as functions of coordinates and time, and if the initial velocity and location of the particle are known, then for a particle moving with a velocity  $v \ll c$  the particle trajectory  $r(t)$  may be determined from the fundamental law of mechanics:

$$m \frac{d^2 r}{dt^2} = f.$$

It is usually mathematically difficult to obtain an exact solution to this problem. An idea of the general nature of motion in a field may be obtained from an examination of the motion of a charge in a uniform field.

**A Particle in an Electric Field.** Assume that a particle enters a field at an angle  $90^\circ + \alpha$  (see Fig. 190). For the choice of coordinates shown in the figure, the equations of motion take the form:

$$\frac{dv_y}{dt} = -\frac{e}{m} E \quad \text{and} \quad \frac{dv_x}{dt} = 0.$$

Whence,

$$v_y = -\frac{e}{m} Et + v_{0y}, \quad v_x = v_{0x}.$$

Integrating again, and assuming  $x = 0$  for  $t = 0$ , we obtain

$$y = -\frac{1}{2} \frac{e}{m} Et^2 + v_{0y}t, \quad x = v_{0x}t.$$

Eliminating time, we obtain an equation of a parabolic curve which describes the motion of the electric charge (dotted in Fig. 190).

If the particle enters the field at right angles ( $v_{0y} = 0$ ), its path is described by the equation

$$y = -\frac{1}{2} \frac{e}{m} E \frac{x^2}{v_0^2}.$$

If the particle enters the field along a line of force, it will continue to move along the line of force with an acceleration  $\frac{e}{m} E$ .

Designating the potential difference between the initial and final positions of the charged particle by  $V$  and using the kinetic energy equation, we obtain

$$eV = \frac{m}{2} (v^2 - v_0^2).$$

If the final velocity  $v \gg v_0$ , then

$$eV = \frac{mv^2}{2} \quad \text{and} \quad v = \sqrt{2 \frac{e}{m} V}.$$

This equation helps to make clear why the unit of energy known as *the electron volt* is widely used:

$$1 \text{ eV} = 1.63 \times 10^{-12} \text{ erg.}$$

An electron volt is the work done in moving an electron through a potential difference of 1 V. This unit may be conveniently employed when the energy refers to a single elementary particle. The work of ionisation and the dislodging and escaping of an electron from a metal range from several to several tens of electron volts.

**A Particle in a Magnetic Field.** The properties of the force acting on a charged particle in a magnetic field are well known (see p. 201).

Assume a particle enters the field with an initial velocity  $v_0$ . Let us resolve this vector into the components  $v_{||}$  and  $v_{\perp}$ , which are parallel and perpendicular to the field, respectively. Then, for motion in the plane perpendicular to the field, we obtain

$$ma = \frac{e}{c} v_{\perp} B.$$

In the longitudinal direction, the particle will move uniformly with a constant velocity  $v_{||}$ .

Motion in the perpendicular plane is circular, and  $a = \frac{v_{\perp}^2}{R}$  is the centripetal acceleration. Thus,

$$\frac{e}{c} v_{\perp} B = \frac{mv_{\perp}^2}{R}.$$

Hence,  $R = \frac{mv_{\perp} c}{eB}$ , i.e. the radius of curvature is directly proportional to the particle velocity and inversely proportional to the magnetic induction. It should be noted that all particles of a given kind in a given field will have the same angular velocity, i.e.,  $\omega = \frac{eB}{mc}$ . Irrespective of the magnitudes and directions of the velocities, the particles will have the same frequency of revolution about a flux line.

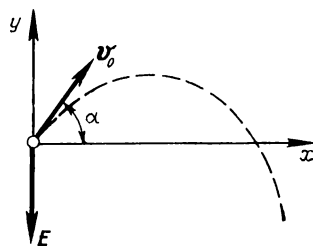


Fig. 190

If a particle enters the field at an incline to the direction of the field, it will move with a frequency  $\omega$  in a spiral of radius  $R$  (Fig. 191). Knowing  $v_{\parallel}$ , the projection of the velocity on the direction of the flux lines, we may determine the pitch of the spiral:

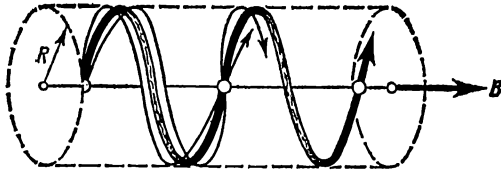


Fig. 191

$$z = v_{\parallel} T = v_{\parallel} \times \frac{2\pi}{\omega} = \frac{2\pi mc}{eB} v_{\parallel}.$$

It is significant that the quantity  $v_{\parallel} = v_0 \cos \alpha$ , where  $\alpha$  is the angle formed between the initial velocity vector and the direction of the field, is constant

to a high degree of accuracy even when the angular spread of the initial velocities is  $5-10^\circ$  ( $v_{\parallel}$  will vary by no more than 1 per cent in such a case). Therefore, every  $z$  centimetres such a divergent beam of charged particles will converge in a point, i.e., focus (within the indicated limits) on a generating line of the cylinder on which the spiral trajectory may be viewed as winding. This generating line passes through the point at which the particle enters the field.

*Example.* Assume that an electron, after being accelerated by a voltage  $V = 300$  volts, enters a magnetic field of flux density  $B = 500$  Gs at an angle  $\alpha = 30^\circ$ . The velocity of the electron is

$$v_0 = \sqrt{\frac{2eV}{m}} = \sqrt{\frac{2 \times 4.8 \times 10^{-10} \times 1}{9 \times 10^{-28}}} = 10^9 \text{ cm/sec.}$$

Note that  $\frac{v_0}{c} = \frac{1}{30}$ . It is pointless, therefore, to introduce a relativistic correction in this calculation.

$$v_{\parallel} = v_0 \cos \alpha = 0.87 \times 10^9 \text{ cm/sec} \quad \text{and} \quad v_{\perp} = 0.5 \times 10^9 \text{ cm/sec.}$$

The radius of the cylinder on which the spiral trajectory of the electron may be viewed as winding is

$$R = \frac{mv_{\perp}c}{eB} = \frac{9 \times 10^{-28} \times 0.5 \times 10^9 \times 3 \times 10^{10}}{4.8 \times 10^{-10} \times 500} = 0.056 \text{ cm,}$$

i.e., its diameter is somewhat greater than a millimetre. The angular velocity is

$$\omega = \frac{eB}{cm} = \frac{4.8 \times 10^{-10} \times 500}{9 \times 10^{-28} \times 3 \times 10^{10}} = 0.89 \times 10^{10} \text{ rad/sec.}$$

The pitch of the spiral trajectory is  $z = v_{\parallel} T = 0.87 \times 10^9 \frac{2 \times 3.14}{0.89 \times 10^{10}} = 0.6 \text{ cm.}$

#### Sec. 171. BEAMS OF CHARGED PARTICLES

In a gas-discharge tube, an electron stream moves in the opposite direction to a stream of positive ions. To obtain an ion ray, i.e., a beam of ions moving in one direction, a hole or canal is made in the cathode. A large proportion of the ions entering this aperture passes through it and then continues to move by inertia. Such beams, called canal or positive rays, were known to physicists as far back as the last century. A similar method of obtaining an ion stream is used even today. First, a substance is transformed into the gaseous state. Then, its molecules are ionised and the positive ions removed from the gas-discharge region through a cathode canal.

A gas discharge is not used to create an electron beam. A so-called electron gun serves as an electron beam source. This is a device based on the phenomenon of thermionic emission (see p. 390). Heated metals, as is well known, may serve as elec-

tron sources. Thus,  $1 \text{ cm}^2$  of tungsten surface heated to  $2,400^\circ$  yields in one second a number of electrons corresponding to a current strength of 1 A.

Fig. 192 is a diagrammatic representation of an electron gun. To accelerate the electrons, a voltage is applied across the electrodes. A tungsten filament (1) heated by an electric current serves as the cathode. The anode (2) has the shape of a glass with a round hole in the bottom. Electrons emerge from this aperture, which determines the divergence and width of the beam. The focussing electrode (3) makes it possible to obtain beams of electrons which are fine and intense (see Sec. 172).

The problem of obtaining an electron beam of maximum intensity for a given expenditure of energy is of great engineering importance.

To utilise all the electrons emitted by the filament, one must, in the first place, accelerate the electrons with a sufficiently high voltage. The filament emits a certain number of electrons per unit time. All these electrons must be drawn away from the filament. If the voltage is low, an electron cloud which impedes emission is formed near the filament. As the voltage is increased, the cloud is gradually dissipated and the thermionic current increases. Finally, we reach a voltage with which no electron cloud is formed. A further increase in voltage does not result in an increase in thermionic current since saturation has been reached. This is the condition required for electron gun operation. Thus, a sufficiently high voltage ensures that all of the electrons are drawn away from the filament.

The next problem is to obtain increased electron emission from a filament. The emission from thorium and oxide cathodes is many times greater than that from tungsten cathodes. A thorium cathode consists of a tungsten wire coated with a very thin layer of metallic thorium. Thoriated tungsten yields the same current at  $1,500^\circ$  as pure tungsten does at  $2,400^\circ$ . An oxide cathode consists of a metallic base coated with a layer of an oxide of an alkaline earth metal. Such a cathode yields the same current at  $900^\circ$  as tungsten does at  $2,400^\circ$ . In modern electronic devices, oxide cathodes are heated indirectly. The cathode is manufactured in the form of a tube in which there is placed a tungsten spiral heated by an electric current.

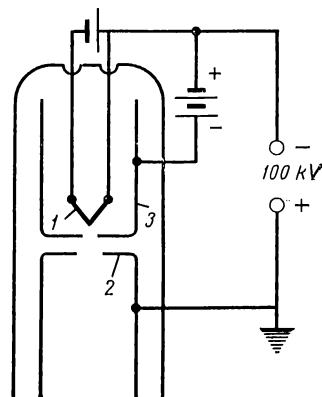


Fig. 192

## Sec. 172. ELECTRON LENSES

An electron beam may be controlled by electric and magnetic fields. The action of such fields is not restricted to the deflection of a beam from its original direction. Thus, a parallel beam of electrons may be made to converge or diverge, and a beam diverging at one point may be made to converge at another. A "lens" for an electron gun is produced by a very simple system of fields. A very important branch of science known as electron optics, the most significant achievement of which is the electron microscope, has developed on the basis of this principle.

Let us recall the properties of an ordinary double convex lens. If an object is placed on one side of such a lens, the image of the object on the other side will be magnified or reduced in size. This is because all rays emerging from an object point gather at an image point and, moreover, all image points are located in a single plane perpendicular to the axis of symmetry of the lens. The simple geo-

metric construction of Fig. 193 shows why the lens operates in such a manner: the angle of deflection of a ray which impinges on a lens is proportional to the distance  $h$  between the axis of symmetry and the point of intersection of the ray and the lens. The construction has been made for an object point lying on the axis of symmetry, but the results are similar for other points as well. It should be stipulated, (as is done in optics) that the discussion is valid if the lens is thin and the beam divergent within the limits of a small solid angle.

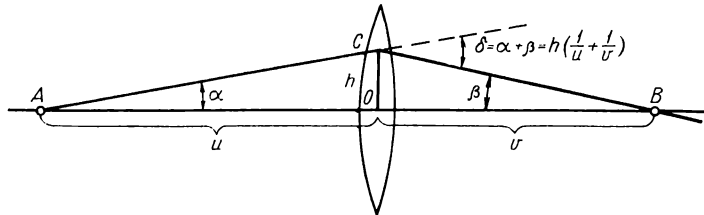


Fig. 193

We shall now show that electric and magnetic fields having axial symmetry may serve as lenses. Such fields may be obtained by means of the following: electrically charged plates with a round aperture in one of them, cylindrical condensers, loops

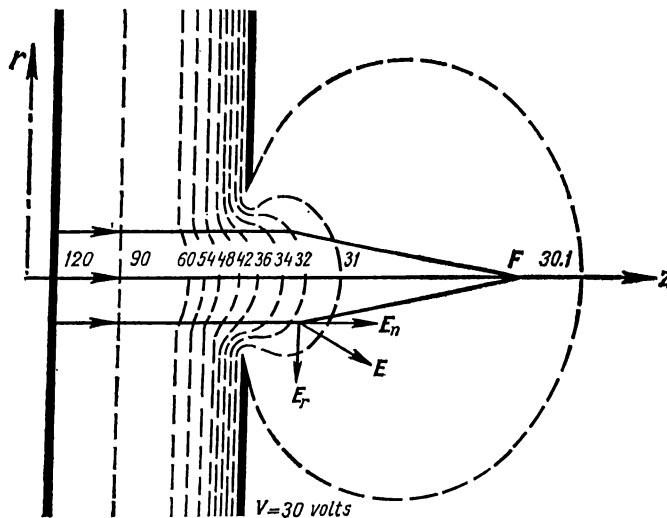


Fig. 194

of current and flat coils. There are a large number of systems which may serve as lenses for electron rays. However, an example of an electrostatic lens and an example of a magnetostatic lens will suffice to clarify the principle.

Let us consider a condenser in which a round aperture has been made in one of the plates (see Fig. 194). If an electron beam impinges on this aperture from the side of uniform field, the beam will be focussed. When an electron reaches the region of nonuniform field, a force perpendicular to the equipotential surfaces, and therefore at an incline to the axis of symmetry, acts on it. Resolving this force into two components, we see that there is a radial component urging the electrons



toward the axis. But this does not suffice for the system to act as a lens. It will also be necessary for the radial component of the field to be proportional to the distance between the axis of symmetry and the point at which the electron reaches the plane of the aperture. It may be easily shown that this is indeed the case. The radial component of the electric field intensity may be expressed in the form

$$E_r = -\frac{1}{2} \frac{dE}{dz} r,$$

where  $\frac{dE}{dz}$  is the field intensity gradient along the axis of symmetry. To prove this, consider a small cylinder oriented as shown in Fig. 195, where the distance 1-2 is infinitely small. Since there is no charge inside the cylinder,  $\pi r^2 dE$ , the flux difference between the ends 1 and 2, must be equal to  $-E_r 2\pi r dz$ , i.e., the flux through the lateral surface with the reverse sign.

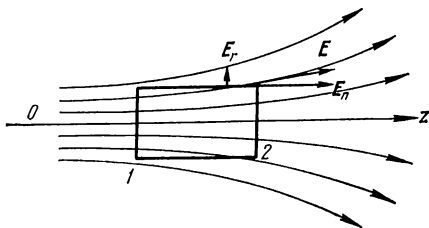


Fig. 195

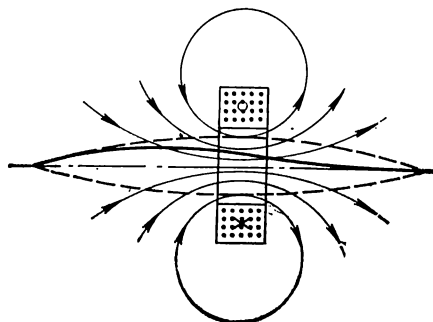


Fig. 196

Thus, an aperture in a charged electric plate serves as a lens for an electron beam.

Now, let us consider the behaviour of electron rays passing through a flat current-carrying coil (see Fig. 196). Such a coil constitutes a magnetostatic lens. The electrons move in a spiral and return to the axis of symmetry after completing one turn of a helix. The focussing properties of the coil are evident. It may be shown that the deflection angle of a ray is proportional to the distance of this ray from the axis of symmetry. The magnetic coil changes the azimuth of the electron trajectory, i.e., in such a lens the image of an object is turned. But this angular displacement does not distort the electron-optical image.

Thus, for an object scattering or radiating electron rays, an "electron image" of the object may be obtained if an electrostatic or magnetostatic lens is placed in the path of the scattered electrons. When a photographic plate or luminous screen is placed in the plane of the image, a peculiar "picture" of the object is obtained. It is bright at points corresponding to the radiation or scattering of many electrons and dark at points corresponding to the absence of radiation or scattering in the object. Since a system of electron lenses yielding a magnified image of an object can be constructed, it is possible to construct an electron microscope.

#### Sec. 173. THE ELECTRON MICROSCOPE

The electron microscope, i.e., a microscope in which the role of a light ray is played by a beam of electrons, provides exceptional opportunities, not yet fully utilised, to "observe" objects directly. This is because the possibilities of magnifying

an object are, generally speaking, unlimited in an electron microscope. On the other hand, an optical microscope provides a magnification of not more than 2,000-3,000.

To understand the reasons for this difference, we must familiarise ourselves with the *resolving power* of a microscope. The question arises: What are the conditions for seeing two close points separately?

Imagine that an ideal point source of light is located in front of a slit or round aperture. When light passes through the aperture, a diffraction pattern is obtained. A lens placed behind the aperture does not concentrate the rays in a point. On the contrary, a blurred circle (or band, in the case of a slit) surrounded by alternately bright and dark rings appears. On p. 278, the magnitude of this blur for a slit was calculated. The radius of the disk, to which a point diffracted from a circular aperture corresponds, was given on p. 279. It is equal to  $1.22 \frac{\lambda f}{D}$ .

Every optical instrument must have an aperture of entry—the objective. Diffraction at the objective is inevitable, and any luminous point in the focal plane of the instrument is diffused into a luminous circle. The angular dimension of the radial blur is equal to  $1.22 \frac{\lambda}{D}$ . Therefore, its linear dimensions in the focal plane are equal to  $1.22 \frac{\lambda f}{D}$ . Here,  $f$  and  $D$  denote the focal distance and the diameter of the objective, respectively. In the case of a microscope, this formula gives merely the order of magnitude since the object is close to the objective and, as a result, the beam of rays cannot be considered parallel. But since we are only interested in the qualitative picture, we shall not go into the fine points.

If two luminous points observed in a microscope are so close that the centres of their luminous image fields are closer to each other than a distance equal to the field radius, these two points cannot be distinguished as separate points.

The limit of linear resolution in a microscope is equal to  $1.22 \frac{\lambda f}{D}$ . Since the ratio of the focal distance to the objective diameter cannot be made significantly less than unity, a microscope enables us to observe two points separated by a distance of the order of a wavelength. Thus, when viewing in ordinary light (wavelength of the order of 0.5  $\mu\text{m}$ ), we cannot detect object details smaller than a hundredth of a micron.

What is the magnitude of useful magnification which may be obtained with an optical microscope? Imagine that a picture is viewed through an ocular, is photographed, then the latter photograph is viewed through an ocular, etc. It is evident that in this manner any desired magnification can be achieved. However, further magnification loses all meaning when it is seen with the naked eye that the resolution limit of points of a photograph has been reached. Thus, if a photograph obtained with an optical microscope is magnified so that 0.5-1 mm corresponds to one micron, the limit of useful magnification has been reached. Hence, the useful magnification of such a microscope is about 1-2 thousand.

As will be shown in the next chapter, an electron ray has the properties of a wave of wavelength

$$\lambda = \frac{h}{mv},$$

where  $h$  is Planck's constant,  $m$  is the mass of an electron and  $v$  is its velocity. When the voltage is equal to 50,000 V, the wavelength equals 0.05 Å. But the distance between atoms is greater than 1 Å. Hence, the usefulness of an electron microscope is not limited by its resolving power.

Calculations indicate that the resolution limit of an electron microscope is  $2\text{--}3\text{ \AA}$ . At present, it is possible to achieve a resolution of  $5\text{--}6\text{ \AA}$ , i.e., a useful magnification of a million.

It turns out that there is much in common between light optics and electron optics. In electron-optical instruments, we find the same elements and the same principles of construction encountered in ordinary optical instruments. The main difference (and this is not of a basic nature) is that the "index of refraction" of an electron-optical lens varies continuously, since the electric and magnetic fields vary

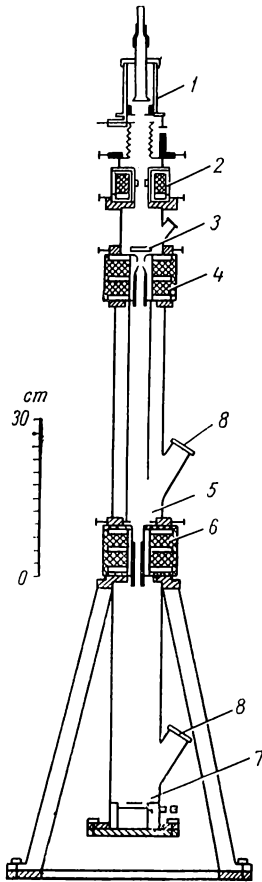


Fig. 197

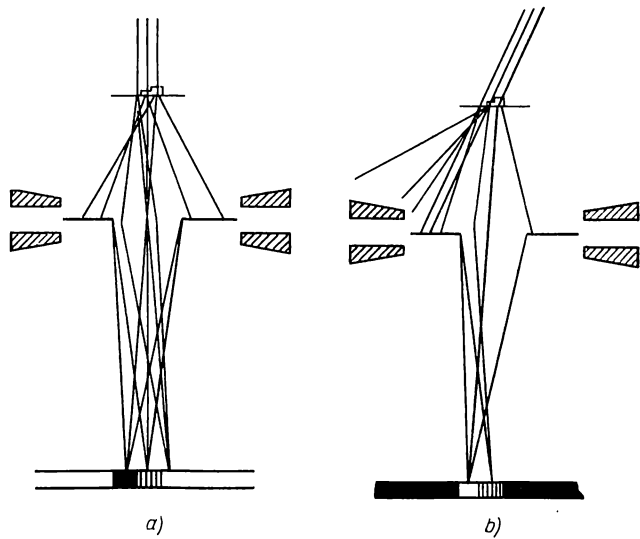


Fig. 198

continuously, while that of an optical lens varies abruptly (at its boundary). Fig. 197 is a diagrammatic representation of an electron microscope: (1) electron projector, (2) condenser lens, (3) object, (4) objective, (5) intermediate image, (6) projection lens, (7) final image, (8) observation window. If we wish to examine an image directly, a fluorescent screen may be used instead of a photographic plate. An electron microscope is much larger than an optical microscope, requires a source of electric voltage and costs considerably more. But this is compensated for by its tremendous resolving power.

The electron microscope portrayed in the diagram employs magnetostatic lenses. A high vacuum, of the order of  $10^{-5}$  mm of Hg, is created in the system in or-

der to prevent electrons from colliding with air molecules. An electron gun produces a beam of electrons with an energy corresponding to 50,000 V. Therefore, the installation must include a high-voltage transformer to boost the line voltage to the indicated value.

Different methods of observing an object by means of an electron beam exist. Since matter is a very strong absorber of electrons, its thickness must be no greater than a fraction of a micron if we wish to observe an object in the "window". When electrons pass through a thin layer of a substance, they are scattered differently by different portions of it. Figure 198 illustrates the two methods used for electron vision. Only those electron rays transmitted through the substance without scattering are allowed to pass, while the scattered rays are blocked by a diaphragm (Fig. 198a). In such a case, the brightest parts of the image correspond to those

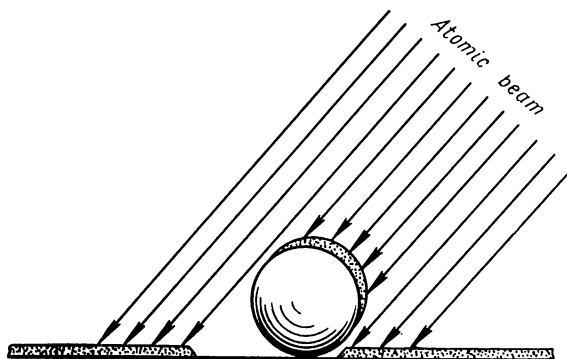


Fig. 199

portions of the substance which do not scatter electrons, including those portions where the layer of substance is particularly thin. On the other hand, the parts of the object which scatter electron rays in profusion are dark. The second method is the reverse of the first (Fig. 198b). The object is placed at an angle to the axis of the microscope, so that only scattered electrons are directed through the lenses. It is evident that the roles of bright and dark fields in the image are now reversed.

The examination of objects in an electron microscope is usually performed on a base the thickness of which is about  $0.01\text{ }\mu\text{m}$ . Such a base is made in the following manner. A drop of a solution of collodion in amyl acetate is placed on the surface of water. The drop spreads on the surface forming a thin film which becomes quite firm after the amyl acetate evaporates. A loop made of thin wire is placed under the film. The object holder is now ready. This base will appear bright for normal incidence of the beam. It will appear dark when the beam impinges at an angle.

If the objects under investigation are poor scatterers of electrons, they will not be seen very well against the common background. The objects are sprayed with a metal to obtain more contrast. The base with the mounted object is placed in the path of a stream of metal atoms produced by vaporising a metal in a vacuum. The spray is directed at an angle to the base, and the sample becomes shaded as shown in Fig. 199. When an object is examined with electron rays an exceedingly bright picture is obtained since electrons are scattered only from the parts of the object sprayed with metal atoms. Fig. 200 shows how flu viruses look under an electron microscope.

The examination of objects on a base is particularly important in biology and medicine. Bacteria are scooped up with the sample holder from the medium in which their presence is suspected. It is easy to study particles obtainable in a suspended state since they can be scooped up with a holder.

Entirely different methods are used in examining the surface of a solid. Under certain circumstances, a solid may be made to emit electrons. By passing the resulting beam of electrons through lenses, we are able to see the surface. However, this procedure cannot always be used: when there is low emission, when the sample

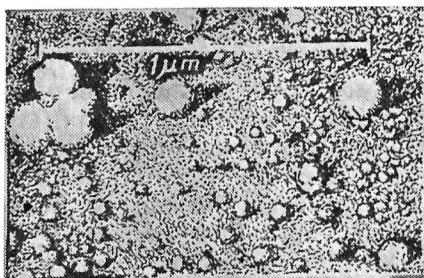


Fig. 200



Fig. 201

cannot be heated, etc. Under such circumstances, the replica method is used. In this method, an object is coated with a thin layer of substance, which may be separated from the object and examined in the opening of an electron microscope. Experiments show that such layers consisting of any one of a variety of substances, e.g., organic, metallic and quartz, form exact replicas of the surface under investigation. A photograph of the surface of frosted glass obtained in this manner is shown in Fig. 201. The replica method requires meticulous experimentation. It is no easy task to separate the layer of substance from the object. One of the methods used is to dissolve the object without damaging the film.

#### Sec. 174. ELECTRON AND ION PROJECTORS

By means of an electron microscope, it has become possible to perceive large molecules as distinct spots or points. But the means are available to achieve considerably more, namely, the shape of a molecule may be discerned and a picture of its electron cloud obtained. This has been accomplished by means of special microprojectors.

Fig. 202 is a diagrammatic representation of an electron and ion microprojector. This consists of a vessel evacuated to  $10^{-8}$  mm of Hg and containing electrodes. The cathode has the shape of a spike the point of which has a very small radius of curvature. It is possible to create near a cathode having this shape a field of the order of  $10^7$  V/cm. For such a field, electrons are torn away from a cold cathode in a radial stream. If an obstacle is located in the path of the stream, a dark image appears on the fluorescent screen (or photographic plate). If an object lies on the surface of the point, the magnification is equal to the ratio of the distance between

the point and the screen to the radius of curvature of the point. Using special means, the radius of curvature may be made less than  $200 \text{ \AA}$ .

If molecules of a substance are placed on the point, their images appear on the screen. This has been done with phthalocyanide molecules, the dimensions of which are about  $15 \text{ \AA}$ . The form of the molecule, its characteristic four-petalled structure,

and the concentration and rarefaction of the electron density were clearly visible on the screen.

Although this method can certainly not be used for all objects under normal laboratory conditions, the possibilities of a method yielding a useful magnification of more than one million should not be underestimated.

But the resolving power may be increased by yet another order of magnitude and, moreover, the clarity of the image may be considerably improved. This may be accomplished by using an ion beam instead of an electron beam, but otherwise employing the

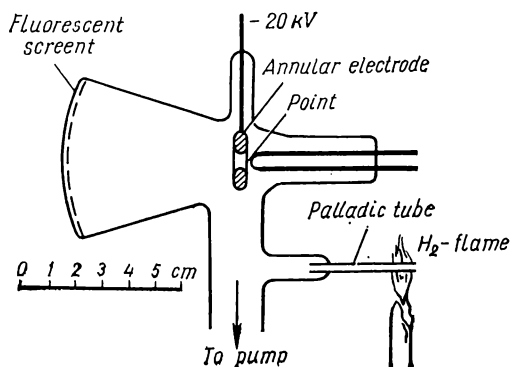


Fig. 202

same principle of object examination. An ion projector does not differ in principle from an electron projector. The point is given a positive potential and when the field is large ( $10^8 \text{ V/cm}$ ) ions may be torn away. For this purpose, it is necessary that atoms or molecules be adsorbed by the surface of the point either beforehand or during operation of the projector. In the instrument shown in Fig. 202, a small quantity of hydrogen molecules is introduced into the vessel by means of a palladic tube. As soon as neutral atoms (or molecules) settle on the surface of the point they give up an electron and then, as positive ions, move toward the screen.

By means of such an ion projector, it has been possible to obtain the image of a tungsten point itself. An image arises owing the fact that adsorption of atoms occurs in specific parts of a tungsten crystal. In the obtained image, it was possible to discern the lattice period, i.e., the resolution attained equalled  $2\text{--}3 \text{ \AA}$ .

#### Sec. 175. THE ELECTRON-BEAM TUBE

An electron-beam tube is a widely used device, being an essential component of a television set, radar system and oscilloscope. The principle of operation of such a tube may be explained by means of the simplified diagram shown in Fig. 203. We see in the figure an electron gun (1) and two condensers (2) for deflecting the electron beam in two mutually perpendicular directions.

Let us consider the application of an electron-beam tube to the recording of rapid processes. If no voltage is applied to the deflecting plates, the electron beam is directed along the axis of the device and a spot appears on the luminescent screen. Assume that an alternating voltage having a frequency greater than  $20 \text{ Hz}$  is applied to the horizontal deflecting plates. Then, the electron beam begins to oscillate in

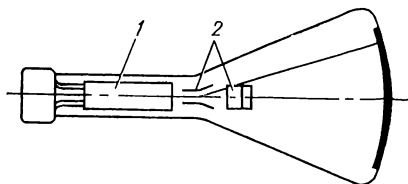


Fig. 203

the vertical direction in synchronism with the varying field. Since electrons have very little mass, these oscillations will have practically no inertia.\* The motion of the beam is not perceived because the luminous spot moves too rapidly for the human eye to follow; moreover the screen has an afterglow.

Now, let us consider the second pair of plates, which provide the so-called "sweep". A saw-tooth voltage is applied across these plates. Since this second pair of plates deflects the beam horizontally, the luminous spot moves, say, from left to right quite uniformly under the action of such a voltage. When the edge of the screen is reached, the luminous spot rapidly returns to its initial position and the process begins anew. By changing the frequency of the saw-tooth voltage within a broad range of frequencies, we can vary the time scale of the horizontal sweep accordingly.

If a sweep voltage is applied to the horizontal deflecting plates and the voltage being investigated is applied to the vertical deflecting plates, a curve of voltage vs. time will be obtained on the screen since the horizontal coordinate of the luminous spot is proportional to the time reckoned from an arbitrary instant.

An oscilloscope is particularly useful in the investigation of periodic processes. It is always possible to select the sweep in such a manner that the curve described by the beam during one run from left to right coincides with the curve described during the second and successive runs. When the sweep period is fixed, we obtain a stationary curve of voltage as a function of time in a given time interval (from a fraction of a period to several periods).

A saw-toothed voltage is produced by a self-sustained oscillatory process analogous to that described on p. 75 (the toppling of a tub of water). The movement of the beam from left to right is produced by continuously and uniformly charging a condenser\*\*. A discharge tube is connected to the terminals of the condenser. As long as the potential difference across the tube is less than the ignition potential, the presence of the tube does not affect the charge on the condenser. When the potential reaches a critical value, the condenser rapidly discharges and the process begins anew. The saw-toothed oscillations must be synchronised with the periodic process under investigation.

An electron-beam tube becomes more complex when a modulator is placed between the cathode and the anode. Such a modulator consists of a metallic cylinder, one end of which is covered with a diaphragm containing an aperture equal to the size of the cathode. A negative potential applied to the modulator makes it possible to control the intensity of the beam. At a certain value of voltage (the blanking voltage), the beam is completely cut off. Such blanking is necessary, for example, during the return trace of the beam. Thus, by means of the modulator, the trace of the beam is blanked during the return sweep.

Two variable quantities may be viewed simultaneously on a screen if an electron-beam tube is equipped with an electron switch that alternately connects the deflection mechanism in one measuring circuit and then in another. Double-beam oscilloscopes have been developed for this same purpose. Such an instrument is equipped with an electron-beam tube having two independent electron projectors and deflecting systems. A double-beam oscilloscope has in addition two separate amplifiers for the voltages being investigated and two saw-toothed voltage generators.

---

\* This absence of inertia is determined by the axial velocity of the electrons. Therefore, to record very rapid processes, high voltage oscilloscopes are used.

\*\* A condenser charges and discharges exponentially, but by using a small portion of the exponential curve these processes may be made quite linear.

The choice of a proper luminescent screen for an electron-beam tube is of great importance. For certain purposes, long-persistent screens are desirable, while for others it is required that the luminosity disappear as soon as the beam is switched off.

Single-pulse processes may be recorded with an electronic oscilloscope if it is equipped with an auxiliary camera the shutter of which is synchronised with the sweep. This makes it possible to photograph the screen at the required instant.

#### Sec. 176. MASS SPECTROGRAPH

The fundamental equation of motion of a charged particle,

$$m \frac{d^2 \mathbf{r}}{dt^2} = e \left( \mathbf{E} + \frac{1}{c} [\mathbf{v} \mathbf{B}] \right),$$

shows that the path of a charged particle is determined by  $\frac{e}{m}$ , the ratio of the charge of the particle to its mass. Therefore, measurements of the deflection of a charged particle in an electric and a magnetic field may be used to find  $\frac{e}{m}$ . Since the initial velocity of a particle is not known,  $\frac{e}{m}$  cannot be determined by measuring the deflection in either an electric or magnetic field alone. The general formulas for deflection in electric and magnetic fields (Sec. 170) show that the path

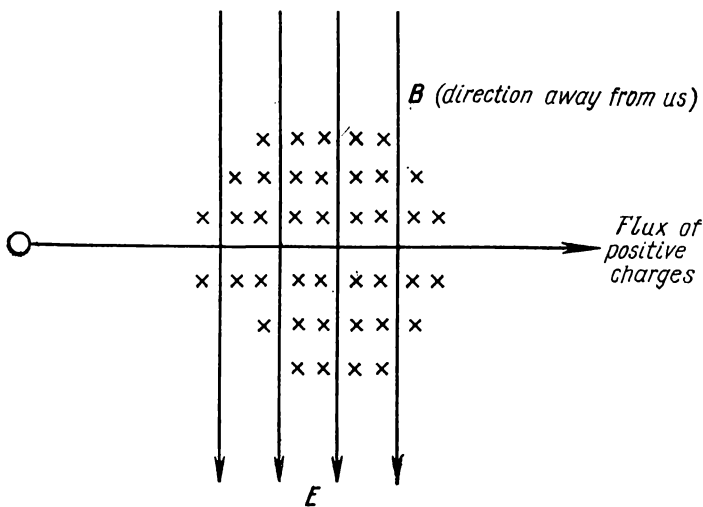


Fig. 204

is determined by coefficients containing  $\frac{e}{m}$  and the initial velocity. The problem is solved by measuring the deflection of one and the same particle in an electric and a magnetic field.

In the simplest case, it is sufficient to balance the electric and magnetic deflections. For this purpose, the fields should be oriented as shown in Fig. 204. Charged particles will not be deflected when the following condition is satisfied:  $eE = \frac{1}{c} evH$ . This experiment enables us to determine the velocity of a particle. Now, it is merely necessary to measure the deflection produced by the electric field or



by the magnetic field alone. Knowing the magnitude of the deflection of a particle from its rectilinear path, we may calculate  $\frac{e}{m}$ .

Measurements of  $\frac{e}{m}$  are of great importance in atomic physics as a means of determining the mass of a particle when the charge is known. This pertains particularly to the determination of the mass of ions.

An instrument in which the particles of a beam may be sorted according to mass, and the composition of the beam according to mass may be investigated, is called a *mass spectrograph*. The mass spectrograph proposed by Aston is represented schematically in Fig. 205. Its principle of operation may be explained as follows. Particles of various velocities are introduced into the electric field of a condenser. Consider a group of such particles having the same  $\frac{e}{m}$  ratio. Upon entering the electric field, a stream of these particles will be divided since fast particles are deflected

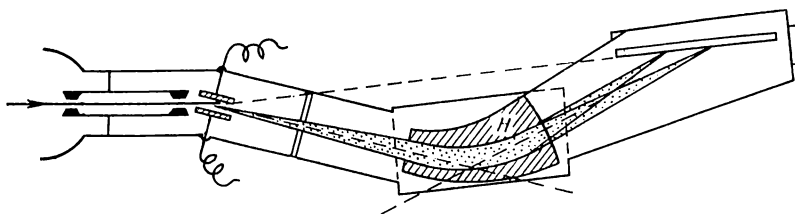


Fig. 205

less than slow ones in an electric field. Now, this spread of particles is introduced into a magnetic field (perpendicular to the page). The sense of the field is such that the direction of deflection of the particles is opposite to that in the electric field. Here, too, fast particles are deflected less than slow ones. Hence, it follows that at some point beyond the field the divided beam of particles will again gather at a point, i.e., become focussed.

Particles having a different  $\frac{e}{m}$  ratio will also gather at a point, but not at the same one. Calculations indicate that the foci of all  $\frac{e}{m}$  lie approximately on a straight line. If a photographic plate is placed along this line, each group of particles will be represented by a separate line.

If a mass spectrograph is constructed with great accuracy, its resolving power will be extremely high and it can be employed to detect the presence of very close isotopes. At first glance, such precision may appear unessential since it may be reasoned that the masses of isotopes differ by at least one atomic weight unit. But while it is true that isotopes of one and the same chemical element differ by an atomic weight unit, isotopes of different elements (e.g.,  $\text{S}^{36}$  and  $\text{Ar}^{36}$ ) may differ very little in mass. Moreover, it is important to be able to determine the mass of complex ions. Such problems arise, for example, in connection with the chemical analysis of gas mixtures. Different particles may then turn out to be close in mass, e.g.,  $\text{C}^{12}\text{H}_2$  and  $\text{N}^{14}$  or  $\text{N}^{14}\text{H}_2$  and  $\text{O}^{16}$ . All such problems may be solved by means of a mass spectrograph.

## Sec. 177. ACCELERATORS OF CHARGED PARTICLES

Actually, all such devices as electron tubes, X-ray tubes and electron guns are accelerators of charged particles, but this term generally denotes installations producing streams of charged particles (electrons, protons, deuterons, etc.) moving with velocities close to the velocity of light. Such streams of particles are then allowed to impinge on matter. The interaction achieved may be used for various purposes: investigation of nuclear transformations, production of radioactive isotopes, medical purposes, chemical action, etc. The role of accelerators in modern science is an extremely important one.

Of course, we can accelerate a particle to any energies, by making it pass in succession through the accelerating fields. But to create particles with energies of tens of thousands of electron-volts, path segments of the order of many centimetres are needed. Modern physics strives for obtaining particle fluxes with energies of tens of milliards of electron-volts. A linear accelerator needed for this purpose would have a length of tens of kilometers. A linear accelerator of enormous length is built in Stanford (USA). Despite some merits of this accelerator, such a solution cannot be regarded as optimal.

Lawrence was the founder of high-energy accelerators. His basic idea is that in a single installation particles should be accelerated by an electric field and repeatedly made to return to the same accelerating gap by means of a magnetic field. The first accelerators operating on this principle became known as *cyclotrons*.

A cyclotron is represented diagrammatically in Fig. 206. The accelerating chamber may be pictured as a flat circular pill box cut along a diameter. It is of large

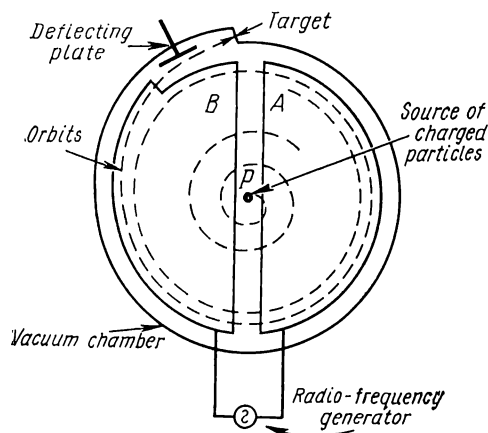


Fig. 206

dimensions, made of metal, and its two halves—the basis of the accelerating chamber—are known as Dees. An alternating electric field of period  $T$  is applied to the Dees. This entire system is placed between the poles of an electromagnet, which creates a strong constant magnetic field inside the box perpendicular to its base.

The magnetic field intensity is determined by the period of the electric voltage. That is, the value of the field intensity must be such that the period of revolution in this field (expressed by the formula  $T = \frac{2\pi mc}{He}$ ) is equal to the period of the electric field. When this condition is satisfied, charged particles entering the accelerating chamber are captured by the fields and accelerated spirally with constant period  $T$ . Thus, a particle in the gap between the Dees is accelerated, traverses half a circle in the magnetic field and arrives at the accelerating gap just when the voltage phase changes by  $180^\circ$ . Thereupon, the particle is accelerated in the same direction and enters the other Dee where it traverses half a circle of larger radius. Each time the particle passes through the gap its velocity increases. The particle must be deflected from its circular path to eject it from the cyclotron.

A cyclotron has limited usefulness. As the velocity of a particle increases its mass increases and, hence, its period of revolution in the magnetic field also increases. As a result, the particle begins to lag, i.e., it arrives at the accelerating

gap when the phase of the voltage has changed by more than  $180^\circ$ . This lag increases until, finally, the electric field not only does not accelerate particles, but retards them. Calculations indicate that the maximum energy that a cyclotron can transmit to a charged particle is given by the formula  $2\sqrt{\frac{ev_0 mc^2}{\pi}}$ . For protons this amounts to 22 million electron-volts (22 MeV), and for  $\alpha$ -particles about three times more. New means are necessary to attain higher energies.

#### Sec. 178. PHASE STABILITY

Veksler in the Soviet Union and McMillan in the United States elaborated a new concept in the accelerator field in 1944 and 1945, respectively. This concept may be summarised as follows: Examination of the formula for the period of revolution of a particle shows that increased mass may be compensated for by an increase in magnetic field. If such compensation takes place, the period of revolution of a charge will remain constant.

Let us assume that the magnetic field intensity increases during the operation of the cyclotron. Among the thousands of millions of particles moving in the accelerating chamber, there are undoubtedly some particles whose increase in mass due to increased velocity is exactly proportional to the increase in magnetic field. Hence, the period of revolution of such particles will remain unchanged. Calculations indicate that under such conditions other particles will also not fall out of synchronism. The only difference, which is quite unimportant in practice, is that the energy of these particles will not increase in a monotone manner, like the energy of the "lucky" particles whose increase in mass is completely compensated for by the increase in magnetic field, but will fluctuate about the energy value of the "lucky" particles.

The orbit radius of the "lucky" particles agrees with their energy. This is the reason for their "good fortune". Now, let us consider a particle the energy of which is greater than that required for a given radius. Such a particle will be retarded owing to an excess of mass increment. On the other hand, if a particle has less energy than required for a given radius, the mass increment will be insufficient and the particle will be accelerated relative to other particles at the same radius. Thus, since the mass of a particle increases with velocity, particles can, so to speak, regulate their own velocity and select voltage phases which serve to correct their motion. That is why this phenomenon is referred to as "phase stability".

It transpires, therefore, that it is possible, in principle, to increase the velocity of particles in a cyclotron without limit if the magnetic field intensity is gradually increased. In such an installation particles must be accelerated in pulses. When the field increases particles are accelerated, but the reverse cycle is idle.

The above method is not the only way to achieve phase stability. Another approach is to slowly vary the period of the electric voltage. The principle is basically the same: an increase in the mass of a charged particle results in an increase in the period of revolution in the magnetic field; in this case, the regime of the alternating electric field is varied so as to compensate for this increase. An accelerator in which the period of the electric voltage is slowly increased is known as a *synchrocyclotron*. The path of a particle in a synchrocyclotron is a flat spiral. The farther this spiral extends, the greater the energy of the particle. Thus, an increase in energy is associated with an increase in the area of the accelerating chamber in the magnetic field. The most powerful accelerator of this type is the synchrocyclotron of the U.S.S.R. Academy of Sciences. This accelerator yields a beam of 680 MeV protons. Its magnet weighs 7,000 tons. The energy limit of a synchrocyclotron is of the

order of hundreds of MeV since a further increase in energy would result in an unthinkable increase in magnet weight. Particle energies of thousands of millions of electron-volts (GeV) are attained by means of proton synchrotrons.

#### Sec. 179. PROTON AND ELECTRON SYNCHROTRONS

Proton synchrotrons are basically different from cyclotrons. Since a proton synchrotron accelerates particles in a single circular orbit, the volume of the magnetic field may be greatly reduced. The entire central region of the magnet is, so to speak, cut out. As a result much less steel is required for the magnet. Thus, the electromagnet of the 10-GeV proton synchrotron of the U.S.S.R. Academy of Sciences weighs 36,000 tons, but the electromagnet of a 10-GeV synchrocyclotron would weigh about 1 million tons.

To accelerate particles in an orbit of constant radius, one must vary the period of the accelerating electric field and the intensity of the magnetic field in a very definite manner. In such an installation, particles are accelerated in pulses. Each pulse is obtained by increasing the magnetic field and decreasing the period of the accelerating electric voltage in a prescribed manner.

For a given orbit radius, a unique relationship exists between the field intensity and the velocity:

$$H = \frac{m_0 v}{\sqrt{1 - \frac{v^2}{c^2}}} \frac{c}{er};$$

and the relationship between the period of revolution and the velocity is determined by the expression

$$T = \frac{2\pi r}{v}.$$

If these conditions are satisfied, a "lucky" particle will be accelerated in a monotone manner.

Since phase stability conditions occur here too, the remaining particles follow a path which oscillates about the circular orbit and they also take part in the synchronous increase in velocity. Since particles oscillate about an average circular orbit, it is necessary to make the width of the pathway for the charged particles rather broad. Methods are being sought which would allow us to reduce the width of this pathway. Success would lead to a reduction in the amount of steel required for a given accelerator and make it practical to construct accelerators with even higher particle energies.

Thus far we have been discussing accelerators of heavy particles. Electron accelerators have a number of special features.

As far back as 1941, an accelerator known as a *betatron* was constructed to accelerate electrons. Such an installation operates on the principle of a transformer. The winding of a magnet constitutes the primary winding and a beam of electrons revolving in a circle of constant radius constitutes the secondary "winding". In other words, the electrons move along a circular line of force of the rotational electric field produced by an alternating magnetic flux.

At first glance, it appears that such acceleration may continue without limit. The increase in mass with velocity does not limit the acceleration since there is no need for synchronism in this phenomenon. Nevertheless, a betatron does have a limit. At energies of several hundred MeV, energy losses due to radiation become considerable—an accelerated electron radiates electromagnetic waves. As a result

of this radiation, the electron path is transformed from a circle to an inward spiral and acceleration becomes impossible. Betatrons are useful when electrons with energies from 20 to 50 MeV are required. If greater energies are desired, *electron synchrotrons* must be used. Such accelerators were first proposed in 1946-47.

The electron synchrotron is similar to the synchrocyclotron described above, i.e., it is a resonance accelerator. An accelerating electric field is added to the magnetic field of the betatron. The accelerating mechanism is maintained by phase stability. But the fact that we are dealing with lighter particles, namely, electrons, simplifies the construction problem. This is because at energies of only 2 to 3 MeV the electron velocity is almost equal to the velocity of light. Hence, when the energy increases further, the radius of the electron path does not change. This makes it possible to construct the magnet in the form of a ring as in the case of the proton synchrotron. The radiation losses of the electron synchrotron are compensated for by increasing the accelerating voltage.

At high-energy levels, radiation losses reach imposing values. In a 300-MeV accelerator having an orbit radius of 1 metre, an electron radiates 1,000 eV per revolution. In a 10-GeV electron synchrotron having an orbit radius of 20 metres, the energy losses per revolution would be equal to 30 MeV.

The capacity of accelerators increases year after year

#### Sec. 180. IONISED GAS

The transformation of an atom into a positive ion, i.e., the removal of an electron from an atom, may be achieved in a variety of ways. Collision with an electron, another atom or a molecule, absorption of a photon—all these methods of transferring energy may result in the ionisation of an atom if, thereby, sufficient energy is transferred to overcome the binding force between the electron and the atom. In the case of different atoms and molecules, this energy ranges from 4 to 25 eV (see p. 390). This means that an atom may be ionised by an electron accelerated by a voltage of 4 to 25 volts. A particle possessing more energy can, of course, transform as many atoms or molecules into ions as its energy reserve permits. One electron accelerated in a powerful accelerator is capable of creating millions of ion pairs.

Ionisation of atoms involves the tearing away of electrons. Ionisation of molecules may also involve the removal of electrons, but in some cases upon ionisation a molecule may break up into two large ions. Thus, ionisation not only creates electrons and positive ions, but may also result in the formation of negative ions. However, negative ions are also obtained by another method, namely, the attachment of a free electron to a neutral atom. It transpires that such a process results in the release of energy (see p. 395).

Ionisation processes in solid bodies will be discussed in detail in Sec. 270. Here, the ionisation process interests us only as a method by means of which streams of ions may be created. From this viewpoint, only the ionisation of gases is of interest. If we wish to obtain a stream of ions of a substance which exists under normal conditions in the solid or liquid state, it is necessary to resort to vaporisation.

Let us consider the ionisation produced in a stream of gas particles. If the source of ionisation is removed, the positively and negatively charged particles will begin to recombine into neutral atoms or molecules. Since upon recombination two particles come together, it is clear that the rate of recombination is proportional to  $n^2$ , the square of the number of ions in a unit volume. If  $\Delta n$  is the number of ions transformed into neutral particles per unit volume in a unit time, then  $\Delta n = -\gamma n^2$ , where  $\gamma$  is a coefficient of the order of  $10^{-6} \text{ cm}^3 \text{ sec}^{-1}$  for most gases under

normal conditions. Under the continuous action of an ioniser, an equilibrium is established between the ionisation and recombination processes. Assume that the ioniser creates  $N$  ion pairs per unit volume in a unit time. At first, the number of ions in the gas increases, but since recombination occurs more and more frequently (proportional to the square of the number of ion pairs present), the increase in ions ceases when  $N = \Delta n$ , i.e.,  $N = \gamma n^2$ . If an ionised gas fills a certain volume, and if the motion of the gas particles is predominantly random, such a conducting gas consisting of neutral and charged particles is called a *plasma* (see below).

The ionosphere is an important example of a highly ionised gas. The number of charged particles in a unit volume of the ionosphere fluctuates considerably, both daily and annually. Such fluctuations, as we know, affect radio reception. There are  $\sim 10^6$  electrons and ions in a cubic centimetre of the ionosphere. The total number of particles in such a volume is  $\sim 10^8$ . Thus, the ionosphere has an ionisation of 1 per cent. In intense plasmas produced in a variety of ionic devices, the degree of ionisation is of the same order of magnitude.

#### Sec. 181. ELECTRIC DISCHARGES IN A GAS

Physicists first studied elementary charged particles by passing an electric current through a gas (electric discharge in a gas). A glass tube filled with gas is incorporated in a circuit and the connections are made to electrodes sealed in the glass. Electric discharges have been studied using various gases, pressures and field intensities.

Different gases generally behave in the same way. A difference in ionisation potential (see p. 390) merely means that certain critical points, to be discussed below, exist at other voltages and pressures.

Let us consider the phenomena, characteristic of every gas, which occur when the voltage applied to a gas-discharge tube is increased. The phenomena to be described take place over a broad range of pressures. We shall merely exclude from our discussion such low pressures for which the free path of the molecules becomes commensurable with the dimensions of the gas-discharge tube and high pressures for which the gas density approaches the density of liquids, i.e., where the molecules do not have a free path. The reader will soon see why the free path of a molecule is of such great importance.

A low voltage is applied across the gas-discharge tube. If there is no ioniser present, current does not flow through the tube. If an ioniser is present, charged particles (ions and electrons) in the gas are urged toward the electrodes by the field. This phenomenon may be called *induced conduction*. The rate at which charged particles migrate toward the electrodes depends on many factors, in particular, the field intensity and gas pressure.

A random thermal motion is superimposed on the ordered motion of the ions and electrons, which occurs under the action of the constant electric force. The particles accelerated by the electric field travel only a short distance, since they inevitably collide with other particles. These collisions are elastic when the velocities are low.

The mean free path of the particles is determined, in the first place, by the gas pressure. The higher the pressure, the shorter the free path and the lower the mean velocity of the ordered motion. On the other hand, the higher the applied voltage, the greater the mean velocity of the ordered motion of the particles.

As indicated in the preceding article, a specific ion concentration is established in a gas under the action of an ioniser; in the equilibrium state, the number of newly formed ions per second is equal to the number of recombinations during this

same interval of time. When a voltage is applied, the equilibrium is disturbed, i.e., some of the ions reach the electrodes before recombination can occur. As the voltage is increased, a larger and larger proportion of the ions created in a unit time reach the electrodes, and the electric current through the gas increases. This tendency continues until no time at all is left for recombination, i.e., all ions created by the ionisers reach the electrodes. It is clear that a further increase in voltage cannot increase the current. The flat portion of the curve in Fig. 207 represents this saturation current.

The lower the gas density, the lower the field intensity which is required to achieve current saturation.

The saturation current is equal to the rate of ion charge formation in the gas-discharge tube.

If the voltage is further increased, new phenomena occur. At a certain point, the electron velocity becomes sufficient to dislodge electrons from neutral atoms and molecules. For this purpose, the voltage across the tube must be high enough to enable an electron to accumulate sufficient energy in its free path to ionise a molecule.

When ionisation occurs as a result of impact, the shape of the current-voltage curve is affected, i.e., the current begins to rise since a voltage increase produces an increase in electron velocity. This higher velocity makes it easier to ionise particles. A large number of ion pairs are created and the current strength increases.

In this voltage range, the passage of current through a gas is accompanied by optical phenomena, i.e., the gas glows. If particle collisions can result in the ionisation of atoms and molecules, all the more so can they result in the excitation of particles, i.e., transitions to a higher energy level. In returning to its normal state, an atom or molecule radiates a quantum of light. We shall not go into this subject here, since radiation by excited atoms and molecules will be discussed on numerous occasions below (see Chapters 28 and 29).

If the electron energy is several times greater than the energy required to ionise a molecule, the passage of electric current through a gas is of a distinctly avalanche nature. When such an electron strikes an atom, an ion and an electron are created. But this new electron is also able to ionise particles. Moreover, the primary electron still possesses sufficient energy to ionise other atoms. The process is cumulative—an avalanche of electric charge, instead of just primary ionisation, moves toward the electrodes. In each successive layer, the number of ion pairs is greater than in the preceding one. When the voltage is fairly high, this avalanche grows extremely rapidly.

The secondary ionisers in the gas are electrons, not ions. The latter are able to ionise gas molecules only when the velocities are very high, i.e., greater than those usually prevailing. If the ions do not produce ionisation, the removal of the external ioniser stops the discharge even when the number of ion pairs due to impacts is hundreds or thousands of times larger than in the primary ionisation. Every avalanche is initiated by a primary electron, and since the electrons migrate toward the anode, the discharge ceases in the absence of an external ioniser as soon as all the electrons arrive at the anode.

Such strong induced discharges possess the following property: for a given voltage, the strength of the electric current passing through the gas is proportional to the number of primary ions created per unit time by the external ioniser. The

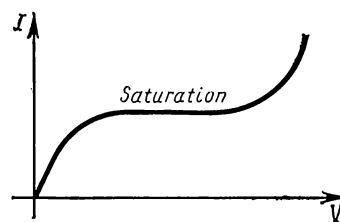


Fig. 207

ratio of the gas-amplified current to the saturation current created by the primary ionisation may reach a value of several thousand. This property of a discharge is utilised in ionisation measuring devices—proportional amplifiers (see p. 422).

An electric discharge may become *self-maintained*, i.e., continue after the external ioniser is removed, only if the ions also become creators of charged particles. This always occurs when the voltage is very high, i.e.—as indicated above—when the ions can ionise upon impact with gas molecules. In such a case, the ions will create more and more new electrons—the primary sources of avalanches.

However, a self-maintained discharge may also occur at considerably lower voltages if the cathode of the gas-discharge tube takes the form of a plate. This is because ions are capable of dislodging electrons from a cold cathode. If the ion velocity is sufficient for such a process, the condition for a self-maintained discharge may be formulated as follows: new electrons appearing at the cathode must, at the very least, replace the work of the primary ioniser.

Thus far we have said nothing about the role of pressure. At high pressures, the discharge column is compressed and thermal ionisation begins. When the pressure changes, the current density distribution changes and so does the luminous nature of the gas discharge. At normal and higher pressures, various kinds of discharges are encountered, e.g., silent, arc and spark discharges. In the rarefied gases, glow discharges occur.

A discharge is said to be silent when a leakage of charge from a condenser or

other charged body is unaccompanied by sound or luminosity. Self-maintained silent discharges—brush discharge and corona—may occur on spikes, thin conductors and, in general, wherever sharp drops in potential, and, therefore, large field intensities, exist.

At higher voltages, spark discharges occur, i.e., the gas breaks down. The breakdown voltage depends almost exclusively on the gas pressure in the region between the electrodes. At normal pressure, the air between spherical electrodes breaks down when the field intensity is 30 kV/cm. Measurement of the breakdown distance may serve as a measure of high voltages.

An electric arc is a special type of discharge. The current density in such a discharge is large even though the voltage between the electrodes is low. The distinctive feature of an arc discharge, usually created between carbon electrodes, is the extraordinary high temperature attained by the electrodes. Therefore, thermionic emission from the cathode plays a large role in an arc.

In the rarefied gases, a glow discharge has a characteristic form for every pressure. The degree of rarefaction may be determined experimentally with great accuracy by mere observation of the discharge form. Various types of gas discharge forms are shown in Fig. 208.

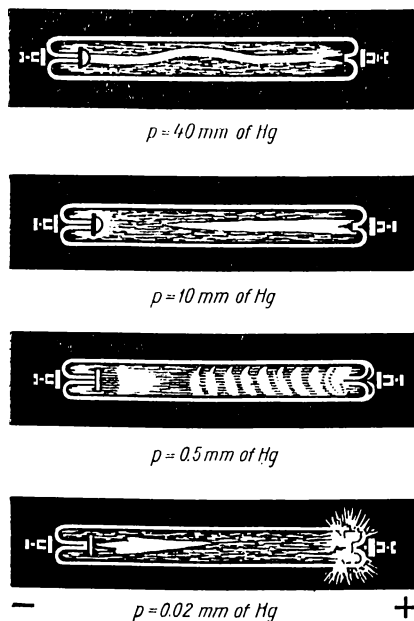


Fig. 208



## Sec. 182. PLASMA

**A substance in the state of plasma.** A gas can be brought in an ionised state by increasing its temperature. Thermal ionisation of gas begins at a temperature of the order of  $6,000^{\circ}\text{C}$ . The average kinetic energy of molecules ( $3/2 kT$ ) becomes already sufficient to ensure frequent collisions among molecules which supply energy necessary for breaking away an electron or for some other kind of ionisation.

The degree of ionisation depends on the gas temperature and gas pressure. With an increase in the gas pressure ionisation decreases.

At temperatures of the order of tens of thousands of degrees and higher a gas of neutral atoms and molecules contained in a certain volume develops into a new state (as distinguished from the gaseous, liquid, and crystalline states) which is called *the plasma*.

It is not difficult to estimate that at temperatures of  $20,000$  to  $30,000^{\circ}\text{C}$  hydrogen gas, for instance, whose density corresponds to a pressure of  $1$  mm Hg at room temperature will turn out to be completely ionised. Indeed, the average energy per degree of freedom at a temperature of  $30,000^{\circ}\text{C}$  is equal to  $1/2 RT = 30$  kcal/mol. This is essentially more than the energy of ionisation of a hydrogen atom.

Thus, thermal collisions will turn a neutral gas into a mixture of two "gases": of the "gas" of protons and the "gas" of electrons. This is just the state of plasma.

A plasma produced from other substances can have a more complicated composition. It may contain electrons, bare nuclei, and various ions. Of course, it contains a certain amount of neutral particles. But at high temperatures this percentage is too low. In the example given above one neutral atom would be for  $2 \times 10^4$  charged protons.

The high-temperature plasma produced by thermal ionization is an equilibrium, or, in other words, isothermal, plasma. Its degree of ionization is very high and it is therefore a very good conductor. The conductivity of high-temperature plasma is comparable to that of metals.

The temperature of the surface of the sun and stars is several thousand degrees, and their interior is heated to millions of degrees. Hence, a considerable part of the matter in the universe is in the state of a high-temperature plasma. The upper layer of atmosphere, the so-called ionosphere, is also a plasma.

It is, of course, impossible to obtain matter in the state of plasma under ordinary terrestrial conditions by heating some vessel because of the absence of proper refractory materials. But with the aid of specially chosen forms of magnetic fields even hot plasma can be kept in a limited volume.

If all particles of a plasma freely exchange their energies, then the plasma will rapidly come to the state of equilibrium, i.e. the average kinetic energy of the electrons and ions will be the same, despite the considerable difference between the masses of the particles. The ions of a plasma move slower than the electrons. In a number of calculations they may be regarded even immovable.

In the case of high-temperature plasma (of the order of  $10^8$  K) the rate of establishing equilibrium among various types of particles varies from smallest portions of a second to several seconds.

A gas-discharge plasma is an example of nonequilibrium plasma. The energy of outside sources is transferred first of all to electrons, and equalization of the energies of electrons and ions will occur only after a large number of collisions. Therefore in a gas discharge the electron temperature  $T_e$  is much greater than the ion temperature  $T_i$ . In an arc discharge  $T_e$  is of the order of many tens of thousands of degrees, whereas  $T_i$  is of the order of thousands of degrees.

To give an idea of how particles behave in plasma, we cite the results of simple calculations from the book "Elementary Plasma Physics" by Academician Lev Artsimovich ("Atomizdat", Moscow).

For the hydrogen plasma of high concentration ( $10^{14}$  ions per cubic centimetre) we obtain for a cold plasma ( $T = 10^4$  K)

$$\lambda \approx 0.03 \text{ cm} \quad \text{and} \quad \tau = 4 \times 10^{-10} \text{ s.}$$

For a hot plasma ( $10^8$  K) the length and time of a free path are respectively equal to

$$\lambda = 3 \times 10^6 \text{ cm} \quad \text{and} \quad \tau = 4 \times 10^{-4} \text{ s.}$$

These results illustrate the collisions of electrons with ions.

Let us now consider the problem of an electric field of plasma. It changes greatly both in space and with time. Nevertheless, we can calculate an average field of a system containing an equal number of ions and electrons situated at a certain average distance  $l$  from one another. It is not difficult to grasp that due to neutrality of the plasma the average field of the plasma (by the order of its magnitude) must be equal to the field of one charge at a distance  $l$  from it, i.e.  $E \approx en^{2/3}$ , where  $n$  is concentration. Thus, for the hydrogen plasma under consideration  $E \approx 4 \times 10^{-10} \times 2 \times 10^9 \approx 1$  CGS unit. This field changes very rapidly. It can reverse its sign within the time of the order of the time of free path and at a distance equal to the distance between the particles.

Neutrality is a necessary property of plasma and is realized very strictly despite a chaotic motion of the electrons. At a great difference between the concentrations  $n_i$  and  $n_e$  the electric field will immediately begin to push out the excessive particles and to attract particles of the opposite sign. Such automatism is realized with a high accuracy (preventing the slightest deviation from neutrality) beginning with small volumes whose radius exceeds  $\sqrt{T_e/n}$ , i.e. for the plasma in question it exceeds  $10^{-5}$  to  $10^{-3}$  cm.

Plasma is a source of electromagnetic waves with wavelengths lying in a wide range. As is known, deceleration of an electron generates a continuous spectrum of electromagnetic waves (the way in which X-rays are produced) with frequencies from zero to  $E_{\max}/h$ , where  $E_{\max}$  is the maximum energy of electron. To estimate the order of the magnitude of the wavelength in deceleration radiation of plasma, we may put  $E = kT_e$ . Then it will turn out that the deceleration radiation of a cold plasma is visible and infrared, whereas that of a hot plasma is X-ray.

An important source of radiation is a recombination of a proton (ion) with an electron. This recombination, obviously, results in emitting a photon of energy equal to the bond energy of particles of opposite signs.

In addition to radiation of the same character for different substances which are in a state of plasma, plasma radiates characteristic line spectra (their origin is described in Secs. 203 and 207), since it contains certain excited atoms and ions.

**Plasma in a magnetic field.** When not in a magnetic field plasma behaves like an ordinary gas. The reason is that plasma is quasi-neutral: even in quite small (but not microscopic) volumes the total charge of the electrons and positive ions equals zero. Therefore, in the absence of an external magnetic field, phenomena in plasma are described by the ordinary equations of fluid dynamics.

If the plasma is in a magnetic field it displays various features due to the action of the magnetic field on the moving charges. The branch of physics dealing with the behaviour of plasma and other conducting liquids (for instance, liquid metals) in a magnetic field is called *magneto hydrodynamics*, and for large Mach numbers, *magnetogas dynamics*. In the present book we cannot devote space to a detailed

discussion of this new and rapidly developing field. We shall consider only some of its major concepts.

(1) Suppose a certain volume of plasma is moving at a velocity  $v$  across the lines of induction of a field  $B$ . Then an induced emf will be set up in this volume just as in any conductor. If the characteristic dimension of the portion of plasma is  $l$ , then  $\mathcal{E} = vBl$ , where  $\mathcal{E}$  is the induced emf. The resistance of this portion of plasma is  $R = \frac{\rho l}{A} \cong \frac{l}{\gamma l^2} = \frac{1}{\gamma l}$ . According to Ohm's law the current induced in the plasma is

$$i_{ind} = \frac{\mathcal{E}}{R} \cong \gamma v B l^2.$$

In accordance with Lenz's law, the induced current interacts with the field so that the force of interaction opposes the movement of the plasma. In addition to ordinary hydrodynamics forces, the plasma is subject to electromagnetic forces as well.

(2) Precise calculation of this interaction is associated with extremely complex mathematics. But the role of the various forces can be assessed by making use of certain dimensionless criteria similar to the Reynolds number. The magnetic Reynolds number  $Re_m$  is the ratio of the magnetic induction of the field set up by the induced currents to the induction of the external magnetic field. Thus

$$Re_m = \frac{B_{ind}}{B}.$$

To evaluate this quantity, we use the fact that  $B_{ind} = \mu_0 H_{ind} = \frac{\mu_0 i_{ind}}{l}$ , where  $l$  is a characteristic dimension of the portion of plasma, and the induced current is expressed by the above equation. Substituting, we have

$$\begin{aligned} B_{ind} &\cong \mu_0 \gamma v B l \\ Re_m &= \mu_0 \gamma v l. \end{aligned}$$

Large magnetic Reynolds numbers are obtained either for a high electrical conductivity of the plasma, or for considerable characteristic dimensions and velocities. The last two are frequently observed on the astronomic scale and are of significance in astrophysics.

(3) At large magnetic Reynolds numbers ( $Re_m \gg 1$ ) the motion of the plasma in the magnetic field should set up a very strong induced magnetic field, many times stronger than the external magnetic field. This requires energy which is made available only at the expense of the kinetic energy of the plasma. Consequently, the induced currents, interacting with the external magnetic field, oppose the motion of the plasma across the field.

At  $Re_m \gg 1$  it may turn out that the plasma practically cannot move with respect to the field. In this case the magnetic field is said to be frozen into the plasma. Here any motion of the plasma is accompanied by corresponding changes in the magnetic field so that the plasma does not cross its lines of induction. Inversely, if the external magnetic field is changed, then, at  $Re_m \gg 1$ , the plasma moves accordingly so as to return the freezing-in condition. This feature is resorted for "pinching" and heating the plasma by a rapidly increasing magnetic field.

(4) The second characteristic criterion in magnetic hydrodynamics is the *Alfven number*  $Al$ , equal to the ratio of the energy density of the magnetic field  $w_m = \frac{B^2}{2\mu_0}$  to the kinetic energy of unit volume of the plasma, i.e. to the density of

the kinetic energy  $w_k = \rho v^2/2$ . Thus

$$Al = \frac{w_m}{w_k} = \frac{B^2}{\mu_0 \rho v^2}.$$

The Alfvén number can also be interpreted as the ratio of the pressure exerted by the magnetic field  $p_m = \frac{B^2}{2\mu_0}$  to the dynamic pressure (or head)  $p_{dyn} = \rho v^2/2$ .

(5) At low magnetic Reynolds numbers the plasma can move with respect to the field. This gives rise to magnetic forces which can be evaluated by Ampère's law ( $F = i l B \sin \alpha$ , where  $\alpha$  is the angle between the conductor and magnetic induction vector). Substituting into this formula the value of the induced current from the equation given in (1), we have

$$F_m = i B l \cong \gamma v B^2 l^3.$$

To assess this force it is compared either to the friction drag  $f \cong \eta l v$  or to the pressure drag  $R \cong \rho v^2 l^2$ . We then obtain two new criteria:

the Stuart number

$$N = \frac{F_m}{R} = \frac{\gamma B^2 l}{\rho v} = Al \times Re_m$$

and the Hartmann number

$$Ha = \sqrt{\frac{F_m}{f}} = Bl \sqrt{\frac{\gamma}{\eta}} = \sqrt{N \times Re}.$$

The significance of these criteria can be understood from the following example. If a liquid flows along a pipe across a magnetic field the type of flow will be only weakly influenced by the field at low Stuart or Hartmann numbers, and the resistance to motion is due chiefly to the viscosity of the liquid. At high Hartmann or Stuart numbers, the viscosity of the liquid becomes a secondary factor and the resistance to motion is primarily due to the interaction between the liquid and the magnetic field.

\* \* \*

When a magnetic field is superimposed on the plasma, the trajectories of charged particles become directed. A free particle moves along a helix wound on the electric vector of the magnetic field. Displacements across the force lines occur only under the action of collisions. At a high temperature and a strong field a charged particle cannot leave the area of the magnetic field.

Under the action of the magnetic field superimposed on the plasma, the latter becomes compressed by electrodynamic pressure. Figure 208a shows the cross section of a plasma column (1 — chamber wall, 2 — vacuum, 3 — plasma). When viewed along the field, the electron trajectories seem to be circular. We may consider that these currents are added into one ring surface current.

According to Sec. 101, at such mutual position of a current and a field there appears a force directed inside the column. According to Sec. 119, the magnitude of the lateral pressure will be equal to the value of the density of electromagnetic energy which in our case is equal to  $H^2/(8\pi)$  (if the field intensity inside the plasma is considered to be reduced to zero by the fields of ring currents). This pressure balances the gas pressure of the plasma which, in the absence of the field, would lead to immediate expansion of the plasma.

The hopes for a long-term confinement of a hot plasma in a concentrated state were linked with the effect of pressure exerted by a magnetic field. The practical significance of such a possibility will become obvious if we get familiar with ther-

monuclear reactions. As it is stated in Sec. 226, temperatures of the order of  $10^8$  K, if they were realized, would lead to creation of a nuclear reactor.

In a powerful gas discharge the electrodynamic force  $\frac{I}{c} [d\mathbf{l}, \mathbf{H}]$  causes a narrow plasma filament torn off the walls of a discharge tube.

The equation  $p = H^2/(8\pi)$  can be rewritten in the following way. Suppose that the ion- and electron temperatures are equal to each other, then

$$p = 2nkT,$$

where  $n$  is the concentration of particles. Assuming that the filament has the shape of cylinder of radius  $r_0$ , and considering that a skin effect takes place, we can write the following formula

$$H = \frac{2I}{cr_0}$$

for the intensity of the magnetic field on the cylinder surface.

Denoting the number of electrons per unit length ( $\pi r_0^2 n$ ) by  $N$ , we get

$$I^2 = 5.5 \times 10^{-14} NT;$$

the formula is written for a current intensity measured in amperes.

If the initial pressure of hydrogen is equal to 0.1 mm Hg, the radius of the tube to 10 cm, and the intensity of the discharge current to  $5 \times 10^5$  A, then the plasma temperature will turn out to be equal to two million degrees.

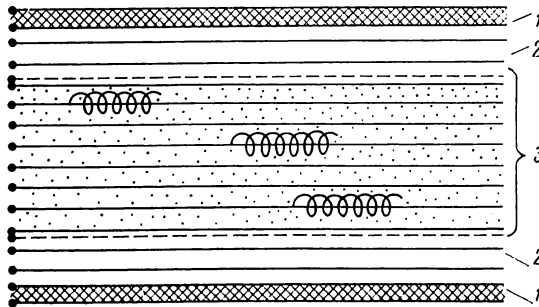


Fig. 208a

**Stability Problems.** Obtaining a stable plasma filament or a plasma formation of other shape is a complicated engineering problem which has not yet been solved. As a result of fluctuations random deformations of the plasma filament can take place.

It seemed at first sight that it is not so difficult to create magnetic "bottles" without leakage. Original theories made it possible to calculate the rates of diffusion in various devices. The results of these calculations were rather optimistic, but the experiment registered the rates of spreading of a plasma column thousand-fold greater than those obtained by theoretical calculations.

During the last two decades the theory of plasma behaviour in a magnetic field has been thoroughly developed, and the causes of the plasma instability have become much more clear. The diagrams, circuits, and patterns of plasma suggested by the original theory which regarded the plasma as a combination of a positively- and a negatively-charged liquids (magneto hydrodynamics) did not present a pre-

cise picture which would demonstrate all complex processes occurring in the plasma. To give an idea of the complications to be necessarily introduced into the theory, let us consider some examples of instabilities which are not taken into account by magneto hydrodynamics.

In a weakly ionised discharge of the type which frequently exists in ordinary fluorescent lamps, when a magnetic field is superimposed parallel to the electric field, the plasma filaments have a spiral form. Up to 1,000-gauss fields the wall diffusion of plasma obeys simple rules. But stronger fields cause intense oscillation of the plasma filament resulting in anomalous diffusion.

Why does it happen? Suppose that one portion of a spiral filament has become denser (marked with a small rectangle on the upper picture of Fig. 208b). The external electric field tends to dissipate this bunch, therefore the ion component of

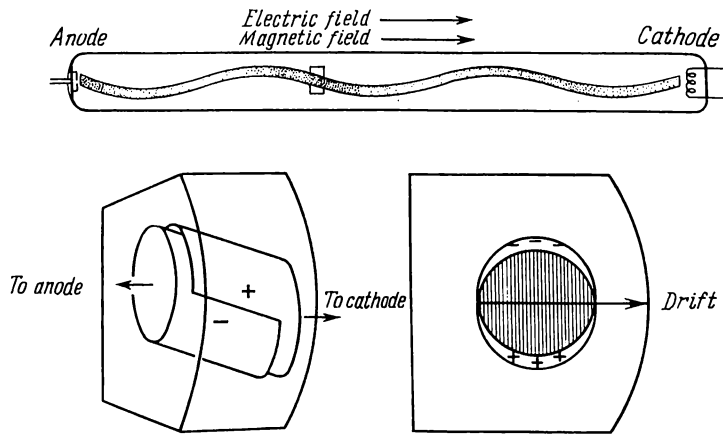


Fig. 208b

this portion of the filament will be displaced to the anode, and the electron component to the cathode (see the left-hand picture of Fig. 208b). Since this portion is of the helical shape, the longitudinal displacement of the negative and positive components will result in the following: in the cross section of this portion of the filament the positive ions will turn out to be displaced with respect to the electrons (see the right-hand picture of Fig. 208b). There will occur a transverse electric field and a corresponding current. But all events take place in the longitudinal magnetic field. Therefore, a Lorentz force will start acting perpendicular to the fields. If we thoroughly consider the geometry of this phenomenon, then it will turn out that the force acts towards the "external" side, i.e. so that the filament gradually spreads to touch finally the walls of the vessel.

And here is another mechanism of instability leading to the leakage of magnetic bottles. Let us assume that in the plasma column there formed a filament with a density exceeding the average (Fig. 208c, the upper picture). Let this region, formed due to fluctuation, be of length  $l$  and exist during time  $t$ . The events we are going to describe now will occur if the thermal velocity of the ions is much less than  $l/t$ , and the thermal velocity of the electrons is much greater than  $l/t$ . It is clear that these conditions are fulfilled readily. Let us show that they lead to spreading of the plasma.

If there is a segment of increased density, then there unavoidably appears a pressure gradient which will begin the process of dissipation. The electrons move

rapidly, and the ions are practically immovable, therefore the middle of the segment gains a positive charge. The electric field thus obtained must eventually balance the pressure gradient.

Let us now consider the cross section of this segment (Fig. 208c, the left-hand picture). It is obvious, that the positive charge concentrated at its centre will be the source of a radial electric field. At each instant, all particles of the plasma move

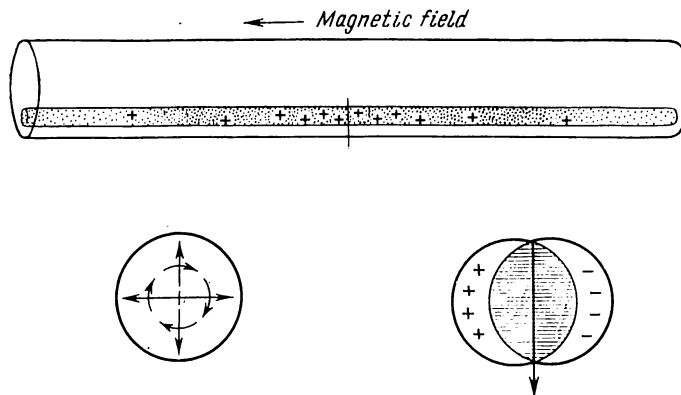


Fig. 208c

in both crossed fields, electric and magnetic. The helical motion about a line of magnetic force is superimposed by a Lorentz displacement across the lines of electric field. In the projection there will appear a circular motion.

The filament instability is due to the fact that in a nonuniform electric field drift velocities for ions are less than for electrons (Fig. 208c, the right-hand picture).

If an electron and an ion (say, proton) are in thermal equilibrium, then their radii of revolution about the lines of magnetic forces  $R_e$  and  $R_p$  will be to each other as 1 to 40. Indeed (see Sec. 170),

$$R = \frac{mvc}{eB}; \quad R_e : R_p = \frac{m_e v_e}{m_p v_p}.$$

But in thermal equilibrium  $m_e v_e^2 = m_p v_p^2$ , i.e.

$$\frac{v_e}{v_p} = \sqrt{\frac{m_p}{m_e}}, \quad \text{hence} \quad \frac{R_e}{R_p} = \sqrt{\frac{m_e}{m_p}} \approx \frac{1}{40}.$$

If the electric field superimposed on the magnetic field is uniform, then the radius  $R$  does not affect the drift velocity. The drift velocity is, of course, proportional to the electric field. If the electric field is inhomogeneous, then a particle moves nonuniformly, i.e. more rapidly within the region of a strong field and more slowly within the zone of a weak field.

Let us now compare the behaviour of an electron and ion moving across the lines of electric force in an inhomogeneous field. Let the axis of a spiral trajectory be projected onto the region where the electric field is the strongest (Fig. 208d). The radius of the ion is forty times the radius of the electron; therefore, in the course of its helical motion about the line of magnetic force the ion will have to "visit" the zone of weak electric fields. Consequently, on the average, the electric force which causes drift will be smaller for the ion.

Let us now come back to the plasma instability due to a random formation of a small dense filament. We showed that in the cross section of this region there occurs a circular drift of particles. Now we know that the ions will be displaced slower. With a uniform density of plasma this circumstance would be insignificant. But with a nonuniform density of plasma the circular drift which turns out to be slow for the ions and rapid for the electrons leads to the following effect: a drift from the regions of high density to the areas of low density results in preferential displacement of electrons, whereas a drift in the opposite direction brings more ions than electrons to the regions of higher density.

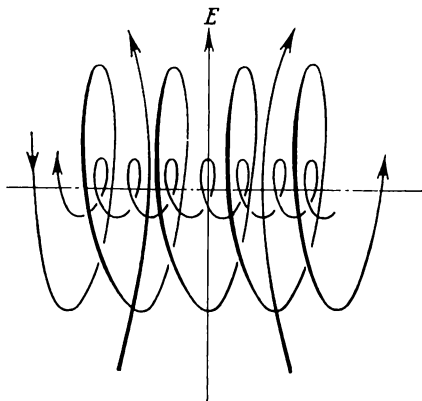


Fig. 208d

The plasma column is necessarily non-uniform in density simply because it has boundaries.

Thus, the density reduces from the axis of the plasma column towards the walls. This means that the drift of the ions and electrons in the region under consideration will lead to separation of the charges in its cross section (see Fig. 208c, right). One more electric field is set in which is shown in the left-hand picture of Fig. 208c. This electric field together with the magnetic field creates a Lorentz force which

acts in the external direction. And so, the dense fluctuation portions are automatically displaced from the axis of the plasma column towards its walls. When it touches a wall, the plasma gives up a part of its energy and begins to cool.

Today, various methods of stabilising high-temperature plasma are being intensively investigated in some of the largest laboratories of the world. Quite a large degree of success has been achieved. But it is still impossible to obtain a plasma with a temperature  $T \geq 10^6$  K, a concentration  $n \geq 10^{20} \text{ m}^{-3}$  and a time of confinement  $\tau \geq 0.1$  s.

Of interest is the fact that some of these parameters have been reached. Thus, in the Soviet device "Ogra-1" a plasma with  $T = 9 \times 10^6$  K and  $\tau = 0.1$  s has been obtained but only for a low concentration:  $n \cong 10^{14} \text{ m}^{-3}$ . In devices with focussed pinch\* (in the USSR and the USA) a plasma with  $T = 2 \times 10^5$  K and  $n = 10^{26} \text{ m}^{-3}$  has been obtained but with a short time of confinement:  $\tau \cong 2 \times 10^{-7}$  s.

Evidently, these instabilities will be overcome in the course of time, and a plasma with the required parameters will be obtained. But when this will come about, and how, are matters that are far from being clear today.

---

\* A rapidly growing magnetic field "pinches" the plasma into a very narrow filament. The process of pinching the plasma takes place so rapidly that the resulting shock wave heats the plasma to a temperature over 10 million degrees, i.e. to the highest temperature ever obtained in the laboratory.



# The Wave Properties of Microparticles

## Sec. 183. DIFFRACTION OF ELECTRONS

Fig. 209 shows X-ray and electron patterns of scattering from the same substance. The close similarity of the patterns indicates that diffraction also occurs in the case of electrons. If the wavelength  $\lambda$  is known, one may determine from a roentgenogram the values of the interplanar distances by means of the equation  $n\lambda = 2d \sin \theta$  (see p. 293). We can calculate  $\lambda$  by measuring the angle  $\theta$  of all the rings on the electronogram and using the values of  $d$  determined from the roentgenogram. The same value is obtained for each ring. This shows that the pattern is produced by diffraction and that a specific wavelength is associated with a beam of electrons.

In order to obtain such a pattern, one must place a thin film of matter in the path of an electron beam. Electrons are easily absorbed by matter and will not pass through a film the thickness of which is more than about  $10^{-5}$  cm. Electronograms may be obtained by means of an electronograph, an instrument similar to an electron microscope. In fact, one may use an electron microscope to obtain diffraction electronograms. For this purpose, it is merely necessary to remove the lens.

Electron diffraction is used for the same purposes as X-ray diffraction, but electronography has a number of advantages over roentgenography. The main advantage is the short exposure time required. Matter scatters electrons much more effectively than X-rays. An electronogram may be obtained in a period of time measured in seconds, while a roentgenogram requires at least several minutes.

Since electron beams do not penetrate matter to any extent, electronography may be conveniently employed to investigate the structure of surfaces. Electronography may also be used to study the distribution of atoms in crystals, assuming, of course, that the structure is not complex.

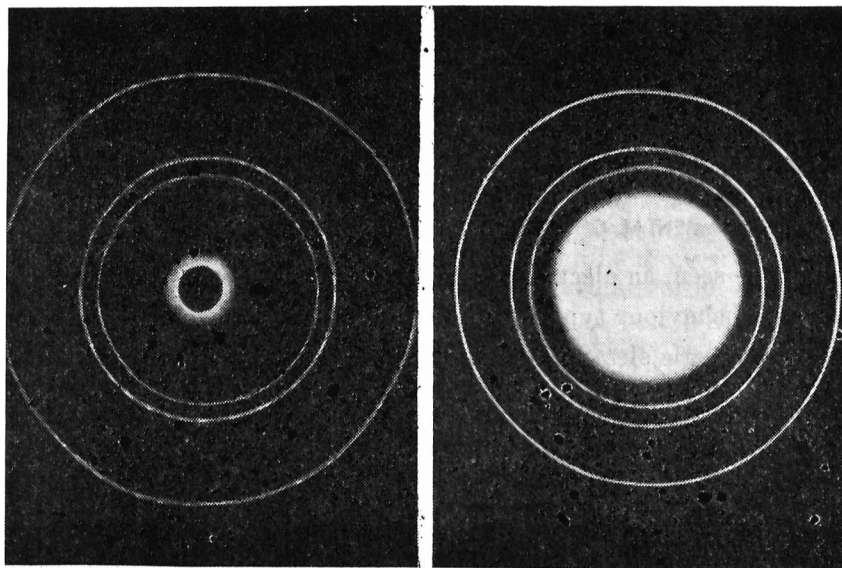


Fig. 209

We are not so much interested in the applications of electronography as in electron diffraction itself. It is of great importance to determine the wavelength of an electron beam. This may be done experimentally. When the accelerating voltage  $U$  is varied, the wavelength varies. It turns out that  $\lambda$  is inversely proportional to  $\sqrt{U}$ . Thus, when  $\lambda$  is expressed in angstroms and  $U$  in volts,

$$\lambda \approx \frac{12.25}{\sqrt{U}}.$$

As far back as 1924, before the discovery of electron diffraction, Louis de Broglie advanced a bold hypothesis. He proposed that the concept of the dual nature of an electromagnetic field, its manifestation in the form of a wave of frequency  $\nu$  and wavelength  $\lambda$  as well as in the form of particles (photons) with an energy  $E = h\nu$  and a momentum  $p = \frac{h\nu}{c} = \frac{h}{\lambda}$ , should be broadened to include particles of matter. Experiments in electron diffraction confirmed this hypothesis. The formula for the electron wavelength

$$\lambda = \frac{h}{p} = \frac{h}{mv}$$

is easily converted into the form

$$\lambda = \frac{\text{const}^*}{\sqrt{U}}, \quad \text{where} \quad \text{const}^* = \frac{h}{\sqrt{2em}},$$

by means of the relation  $\frac{mv^2}{2} = eU$  or  $v = \sqrt{\frac{2}{m}eU}$ .

Substituting the values of  $e$ ,  $m$  and  $h$  in the CGS system of units, and converting  $U$  into volts, we obtain the above value for the constant. This undoubtedly proves the correctness of de Broglie's hypothesis.

By means of the above formula, one can obtain the following electron wavelengths for the indicated values of accelerating potential:

$U$ (volts)	1	100	1,000
$\lambda$ (Å)	12.25	1.22	0.39

Since diffraction from a crystal is possible when the wavelength is of the same order as the period of the crystal lattice ( $\sim 1$  Å), it is seen that electrons the energies of which are of the order of a hundred electron volts may be diffracted from a crystal.

#### Sec. 184. THE FUNDAMENTAL CONCEPTS OF QUANTUM MECHANICS

As we have seen, an electron beam may behave as a wave having a wavelength  $\lambda = \frac{h}{p}$ . Is this behaviour typical of a large group of electrons or are wave properties inherent in single electrons? To answer this question, let us compare the electronograms obtained with a low and a high intensity beam of electrons. In one such experiment, the beam was of such low intensity that the average time interval between two successive passages of an electron through the diffracting system was 30,000 times greater than the time taken by an electron to travel from the filament to the photographic plate. Nevertheless, the diffraction pattern differs in no way from another electronogram obtained with a beam the intensity of which was  $10^7$  times greater. Such experiments show that the wave property is inherent in single electrons. Thus, what was stated on pp. 324-326 with respect to a photon is also valid for an electron. An electron does not behave like a projectile or pellet.

The behaviour of an electron cannot be described by means of Newton's laws of mechanics. The field devoted to the investigation of the behaviour of microparticles is known as *quantum mechanics*.

This dual behaviour is not peculiar to an electron alone. Wave behaviour is characteristic of all microparticles. Thus, it is possible, for example, to observe neutron diffraction. Diffraction of helium atoms and hydrogen molecules have also been observed. As will be seen below, the greater the mass of a particle the more do its wave properties recede into the background. But more about this later. Let us designate by  $\psi$  the amplitude of the wave associated with a microparticle of mass  $m$ . This amplitude, like the amplitude of a wave of any kind, is a function of coordinates. The function  $\psi(x, y, z)$  or "psi function" must satisfy the wave equation (see p. 258):

$$\Delta\psi + \frac{4\pi}{\lambda^2} \psi = 0.$$

Let us substitute in this equation the microparticle wavelength  $\lambda = \frac{h}{mv}$ , expressing the velocity of the particle in terms of its energy. If the particle moves in a field with a potential energy  $U$  and has a total energy  $\mathcal{E}$  then

$$\frac{mv^2}{2} = \mathcal{E} - U \quad \text{and} \quad v = \sqrt{\frac{2}{m}(\mathcal{E} - U)}.$$

Substituting in the wave equation, we obtain

$$\Delta\psi + \frac{8\pi^2m}{h^2}(\mathcal{E} - U)\psi = 0.$$

This equation is called *Schrödinger's equation* and is the fundamental law of quantum mechanics. It should be noted that we have merely performed a substitution in the wave equation, which by no means constitutes a derivation of the fundamental law of quantum mechanics. The substitution should be viewed rather as a conjecture leading to the discovery of this equation.

Much like Newton's fundamental law of motion, the Schrödinger equation is a law of nature encompassing, as we shall presently see, an extensive range of phenomena\*.

This differential equation enables us, in principle, to find at all points in the region under consideration the amplitude  $\psi(x, y, z)$  of the wave associated with a microparticle.

How can the validity of the Schrödinger equation be verified? This may be done by means of a basic confirmation of the theory: the intensity of the  $\psi$ -wave at every point in space, i.e., the quantity  $\psi^2$ , is the probability density\*\* of an electron at this point in space. As for the amplitude of the  $\psi$ -wave, it cannot be determined experimentally, like the field intensity vectors of an electromagnetic wave.

The description of a particle by the  $\psi$ -function is not to be regarded as an incomplete, imperfect method of describing its motion. It would be incorrect to believe that it is solely due to the peculiarities of quantum mechanics that a particle has a probability  $\psi^2(x, y, z)$  of being at a given point in space and that by means of a better theory the path of the particle could be determined and its location indicated with certainty. An exact description of the motion of a particle, i.e., the determination of its path, is not possible because the behaviour of a microparticle

\* We have simplified the picture by not considering the dependence of  $\psi$  on time. The exact Schrödinger equation takes this dependence into account.

\*\* That is, the probability of finding the particle in a small volume divided by the magnitude of this volume.

is completely different from that of a large body. A microparticle is not a particle in the classical sense.

Let us again refer to the experiment with two slits which was used to illustrate the dual behaviour of a photon.

Assume that a beam of electrons impinges on a screen in which two slits are cut. Diffraction occurs. Let us direct our attention to one point on the screen where, say, interference annuls the wave. If only one slit is kept open, electrons reach this point. If both slits are open, however, electrons do not reach this point. This phenomenon cannot be explained by the collective behaviour of the electrons.

If we insist on the electron's behaviour as a classical particle, it must be accepted that an electron passing through one slit "knows" whether the other slit is open or closed, and behaves accordingly. It must be concluded that an electron is not a classical particle. Every electron has wave properties and the concept of an electron path is not fully applicable. Therefore, the question whether the electron passed through one slit or the other when both slits were open is meaningless. Such a question is meaningful only for "ordinary" particles, but not for microparticles.

Does this mean that it is meaningless to speak of the velocity and coordinates of a microparticle? This question is answered by the so-called indeterminacy (or uncertainty) principle formulated by the German physicist Heisenberg.

#### Sec. 185. THE UNCERTAINTY PRINCIPLE

This principle indicates the limits within which the classical description of particles applies to microparticles.

**Applicability of the Path Concept.** Let us assume that we wish to determine the coordinate of a microparticle at some point  $x$  and are able to do this with an accuracy of  $\Delta x$ . To "see" a microparticle, one must use a "microscope" in which the wavelength of light is sufficiently short (the shorter the wavelength the greater the resolving power). In principle,  $\Delta x$  may be made as small as desired. For this purpose, it is merely necessary to reduce the wavelength until  $\Delta x$  is of the same order of magnitude as the wavelength:  $\Delta x \approx \lambda$ .

However, if the wavelength is short, this means that the corresponding photon has a large momentum:  $p = \frac{h}{\lambda}$ . This momentum will be transmitted to the particle being "observed" under the microscope, i.e., when the particle is "flicked" its momentum changes by a  $\Delta p$  the order of magnitude of which is  $\frac{h}{\lambda}$ . By decreasing  $\lambda$ , we decrease  $\Delta x$ , the uncertainty of the coordinate, but at the same time we increase  $\Delta p$ , the uncertainty of the momentum. Eliminating  $\lambda$  from the following relations:  $\Delta x \approx \lambda$  and  $\Delta p \approx \frac{h}{\lambda}$ , we obtain the equation

$$\Delta x \times \Delta p \approx h,$$

the expression for the *uncertainty principle*. This indicates that specifying the path of a microparticle has physical meaning only if it is understood that the coordinates and momentum in a given direction have uncertainties which satisfy the inequality

$$\Delta x \times \Delta p > h.$$

This remarkable relation establishes the limits of applicability of classical physics to a microparticle.

Substituting the product  $mv$  for the momentum  $p$ , which is justified in the case of velocities which are not very close to the velocity of light, we obtain the condi-

tion

$$\Delta x \times \Delta v > \frac{h}{m},$$

a relation between the uncertainties in a coordinate and the corresponding particle velocity along the  $x$ -axis. The right member of the relation will vary by many orders of magnitude, depending on whether we are dealing with an electron, atom, molecule or tennis ball.

For an electron

$$\frac{h}{m} \approx 7 \text{ cm}^2/\text{sec},$$

hence, the uncertainties in a coordinate and the corresponding particle velocity are related as follows:

$$\Delta x \times \Delta v > 7.$$

Consider an electron located within an atom, the diameter of which is  $10^{-8}$  cm. Can the motion of the electron in the atom be described as if the electron were an "ordinary" particle? Using the uncertainty principle, we find that  $\Delta v \approx 10^8$  cm/sec. Thus, we can only speak of the velocity of an atomic electron in very general terms. The concepts on an electron path in an atom, an electron path of transition from one energy state to another (see below), etc., are meaningless. In short, an atomic electron is quite different from an "ordinary" particle.

Now, let us assume that an electron has entered a Wilson cloud chamber and that we wish to trace its path with an accuracy of the order of several tenths of a millimetre. If the width of the particle track is  $\Delta x = 10^{-2}$  cm, then according to the uncertainty principle  $\Delta v \approx 7$  m/sec. This is the uncertainty of the lateral component of the velocity. Even if the electron velocity is only 1 km/sec an uncertainty of the indicated order is insignificant and specifying the electron path becomes meaningful. Similarly, we can speak of the path of an electron beam in a microscope and the path of electrons in an electron-beam tube without coming into conflict with the "classical" picture.

The mass of protons, neutrons, atomic nuclei and atoms is thousands times greater than that of an electron. Therefore, the classical paths of such particles are of somewhat greater usefulness. Thus, for example, in the case of an  $\alpha$ -particle, the mass of which is about 7,000 times greater than the mass of an electron,

$$\Delta x \times \Delta v > 10^{-3}.$$

Is it meaningful to ask at what location in an atom did the path of an  $\alpha$ -particle penetrating a substance pass through the atom? We wish to trace the path with an accuracy of  $10^{-9}$  cm and we have at our disposal data on the lateral component of the velocity the uncertainty of which is  $10^6$  cm/sec = 10 km/sec. For a fast  $\alpha$ -particle (20,000 km/sec), this uncertainty is insignificant. Therefore, it is possible to say whether the path of the  $\alpha$ -particle penetrating the atom passes far from the centre of the atom or not.

On the other hand, it is meaningless to speak of the path protons or neutrons in nuclei, since the size of a nucleus is  $\sim 10^{-13}$  cm.

For large molecules, e.g., proteins having a molecular weight of the order of  $10^6$ , the uncertainty principle is of no significance. Here,  $\Delta x \times \Delta v > 10^{-7}$ ; hence one can reliably define the path of such a molecule in considerable detail. Even the random thermal motion of such a molecule, the average velocity of which is of the order of 1 m/sec, may be traced, and the path of its centre of gravity indicated with an accuracy of the order of 1 Å.

Needless to say, a particle of dust, even if it is visible only under a microscope, is too large for the uncertainty principle to be of practical significance.

**The Simultaneous Measurement of Two Physical Quantities.** It should not be thought that the inability to determine the path of a particle is due to a measuring deficiency which will eventually be overcome by physicists. The lack of meaning in specifying a pair of physical quantities with ideal exactness is a peculiarity of microparticles. The methods of describing the behaviour of an "ordinary" particle are inapplicable to a microparticle. Only for a classical particle is it meaningful simultaneously to specify and define its coordinate and momentum.

The principle of uncertainty has broader significance than being simply a means of judging whether or not the path of a particle can be determined. As an integral part of the mathematical apparatus of quantum mechanics, the principle of uncertainty enables us to evaluate the possibilities of simultaneous measurement of any physical quantities, not only coordinate and momentum.

First, let us define the uncertainty principle in relation to coordinate and momentum. It should be recalled that a particle has three coordinates and that a momentum vector has three components. Instead of the one relation discussed above, three relations should be written, namely:

$$\Delta x \times \Delta p_x > h; \quad \Delta y \times \Delta p_y > h; \quad \Delta z \times \Delta p_z > h.$$

The possibility of simultaneously determining (specifying) all the coordinates along the different axes, and all the momentum components, is also considered in quantum mechanics. It turns out that it is possible (meaningful) simultaneously to specify the coordinates or simultaneously specify all three momentum components. Why is this point being emphasised? It would seem that it is always possible simultaneously to determine the three components of any vector. Careful consideration indicates that this is not so. An example of a vector the three components of which cannot be determined simultaneously is the angular momentum (i.e., moment of momentum) of a particle.

Assume that a particle rotates about an axis with an angular momentum  $L$ . This motion may be viewed as the resultant of three rotations about three mutually perpendicular axes with angular momenta  $L_x$ ,  $L_y$ , and  $L_z$ . In the case of an "ordinary" particle, the three components of the angular momentum may be determined separately since the path of the particle may be traced. In the case of a microparticle such a determination is not possible, and simultaneous specification of all three components of the angular momentum is meaningless. To clarify this point, let us assume the converse for a moment, namely, that all three components of the angular momentum are known. But then the total angular momentum could be constructed from the three components. In that case, the plane in which the particle moves is determined. But if this plane is known, then we know precisely the coordinate of the particle along the axis of rotation and note simultaneously that the velocity of translatory motion along the axis of rotation is equal to zero. This contradicts the principle of uncertainty relating coordinate and momentum.

Thus, it is characteristic of a microparticle that it is not possible simultaneously to determine the three components of its angular momentum. What data relative to its rotation may be specified simultaneously? The uncertainty principle gives the following answer: any component and the absolute value (vector length) of the angular momentum. There is one exception to this rule: complete absence of rotation may be established for a microparticle, i.e., the angular momentum vector may equal zero; in other words, all three components equal zero simultaneously.

**Energy and Time Interval.** On the basis of the uncertainty principle relating the coordinate and momentum of a particle, one may suspect that a more or less analogous relation which involves energy exists. Thus, in ordinary particle mechanics, it was necessary to know simultaneously the position and velocity of a particle in order to calculate the total energy as the sum of potential and kinetic energy. For a microparticle this is not possible. However, the total energy of a particle may be found as a whole, i.e., without separation into parts, and on the basis of what was just said it is natural to expect that this may be done with an uncertainty  $\Delta\mathcal{E}$ . If it is assumed that the uncertainty principle maintains its form, then from dimensionality considerations we must conclude that the uncertainty relation for energy must have the following form:

$$\Delta\mathcal{E} \Delta\tau \approx h$$

where  $\Delta\tau$  is the time interval.

What is the significance of this time interval and how should the uncertainty relation for energy be interpreted? The time interval  $\Delta\tau$  is the time during which a microparticle possesses an energy  $\mathcal{E} \pm \Delta\mathcal{E}$ . The uncertainty in the energy of a microparticle is determined by the time during which it is in the given energy state. The uncertainty in energy becomes significant only when the time during which it is at the given energy level is measured in minute fractions of a second.

An atomic electron remains for an indefinitely long period of time at its lowest, or fundamental, energy level (see below for a detailed discussion). Therefore, the energy of the fundamental state is fixed quite rigidly. An electron remains for a very short period of time at a higher level. Its energy in such a state is  $\mathcal{E} \pm \Delta\mathcal{E}$ . Accordingly, when an atom passes from a higher energy level to a lower one, the radiation frequency cannot be exactly defined, i.e., it lies in the narrow band  $\nu \pm \frac{\Delta\mathcal{E}}{h}$ . This may be observed experimentally: spectral lines are of finite width, which may be used to determine the so-called lifetime of a microsystem in an excited state. Experiments show, for example, that the width of spectral lines in the X-ray region is of the order of 10 eV. Thus, in such a case, the lifetime in an excited state is of the order of  $\frac{h}{\Delta\mathcal{E}} \sim 10^{-16}$  sec.

#### Sec. 186. THE POTENTIAL SQUARE WELL

On page 42, we discussed potential curves, which clearly illustrate the conditions of particle motion. The simplest curve of this type is a right-angled curve, a so-called *square well*. In such a well, the potential energy has a constant value over a segment  $a$  (for simplicity, we restrict ourselves to the linear case). At the edges of the well, the potential energy changes abruptly. If the walls are very high, the potential energy inside the well may be considered equal to zero (since the choice of origin is immaterial), and at the edges of the well equal to infinity (see Fig. 240).

Assume that an electron (or some other particle) is located in the well. Let us try to determine the nature of its motion for the simple one-dimensional case, i.e., let us assume that the electron is moving along the  $x$ -axis. If Newton's laws of mechanics were applicable to an electron, such an electron would move continuously — first toward one side of the well, where it would be elastically reflected from the wall, then toward the other side, etc. There is no other possibility from the viewpoint of Newtonian mechanics, since for  $U = 0$  the kinetic energy  $\frac{mv^2}{2}$  is constant. Thus, for motion in a square well, the following conclusion may be

drawn from the mechanics of "ordinary" particles: A particle may move in the well with any kinetic energy  $\frac{mv^2}{2}$  or it may remain motionless. For any given energy, the motion is uniform—first toward one side, then toward the other, i.e., the velocity reverses direction abruptly at the end of the allowed interval.

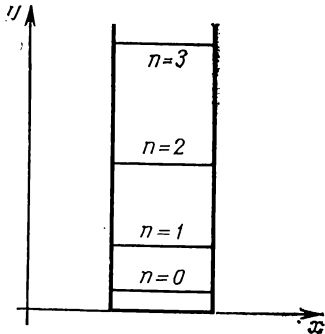


Fig. 210

Now, let us consider the electron motion in such a well from the viewpoint of quantum mechanics.

The behaviour of the electron is characterised by the  $\psi$ -function. The square of this function indicates the probability of finding the electron at some point in the given interval. Since  $U = 0$  inside the well, the Schrödinger equation becomes simplified and may be written in the form

$$\frac{d^2\psi}{dx^2} + \frac{4\pi^2}{\lambda^2} \psi = 0, \quad \text{where} \quad \lambda = \frac{h}{\sqrt{2mE}}.$$

This equation is satisfied by the sine and cosine of the argument  $\frac{2\pi x}{\lambda}$ . If the well is bounded by the coordinates  $x = 0$  and  $x = a$ , then for these values  $\psi = 0$ , since the electron does not penetrate the walls. Therefore,  $\cos 2\pi \frac{x}{\lambda}$  is not suitable as a solution to the equation ( $\cos 2\pi \frac{x}{\lambda} = 1$  at  $x = 0$ ). Hence,

$$\psi = A \sin \frac{2\pi}{\lambda} x.$$

But the wavelength  $\lambda$  cannot be arbitrary, since  $\psi = 0$  at  $x = a$ . The following equation must, therefore, be satisfied:

$$\frac{2\pi}{\lambda} a = (n+1) \pi$$

or

$$\lambda = \frac{2a}{n+1}, \quad \text{where} \quad n = 0, 1, 2, \dots$$

Thus, the wavelength may assume the values  $2a, a, \frac{2a}{3}, \frac{a}{2}, \dots$ . It is seen that the  $\psi$ -function represents the amplitude of a standing wave (see p. 98) and that formally this problem has much in common with that of a vibrating rod or string. But if the wavelength  $\lambda$  has a discrete set of values, then the energy  $\mathcal{E}$  of a micro-particle cannot be arbitrary, i.e.,

$$\mathcal{E} = \frac{(n+1)^2 h^2}{8ma^2};$$

the integer  $n$  is called the *quantum number*.

Thus, the Schrödinger equation leads to the quantisation of energy. A micro-particle in a square well has a discrete set of energy levels. The lowest energy level occurs for  $n = 0$ . This energy is equal to  $\frac{h^2}{8ma^2}$  and is the zero-point energy of a particle located in a square well.

The existence of a zero-point energy is an interesting peculiarity of microparticles. In the case of "ordinary" particles, the lowest energy has a value of zero. But



under no circumstances can microparticles cease to move. At absolute-zero temperature, a microparticle has a specific zero-point energy, the values of which differ considerably depending on the nature of the force field in which the particle is located.

*Example.* Assume that  $a = 1 \text{ \AA}$  (a characteristic value for an atomic region). Then, the zero-point energy of an electron in the square well is

$$\mathcal{E}_0 = \frac{h^2}{8ma^2} = \frac{(6.6 \times 10^{-27})^2}{8 \times 9.1 \times 10^{-28} (10^{-8})^2} = 0.6 \times 10^{-10} \text{ erg} = 37 \text{ eV}.$$

If  $a = 1 \text{ cm}$  (a free electron in a piece of metal),  $\mathcal{E}_0 = 0.6 \times 10^{-26} \text{ erg} = 37 \times 10^{-16} \text{ eV}$ .

The velocity of an electron at a given energy level may be calculated by means of the wavelength:  $v = \frac{h}{m\lambda}$ . But in such a case, the electron motion cannot be described by the equations of classical mechanics. It is not possible to indicate

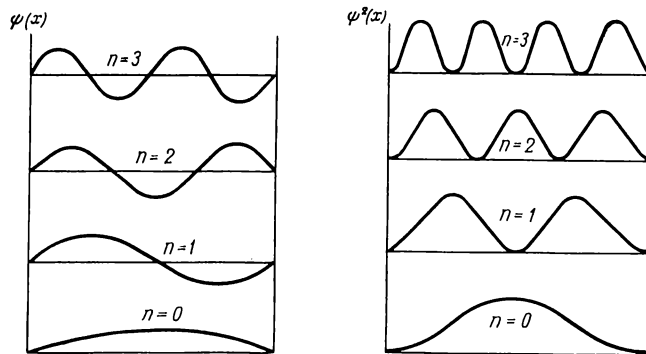


Fig. 211

where the electron is located at one or another instant of time. However, the quantity  $\psi^2$ , i.e., the probability density of the electron at one or another point in space, may be determined from the equation

$$\psi_n^2 = A^2 \sin^2 \frac{2\pi}{\lambda} x = A^2 \sin^2 \frac{\pi}{a} (n+1) x.$$

It should be noted that to each energy level there corresponds a proper (of same number  $n$ ) wave function (also called characteristic function or eigenfunction).

Fig. 211 shows the  $\psi$ -function and  $\psi^2$ -function for the first four energy levels of an electron located in a square well. Quantum mechanics leads to the conclusion that an electron does not appear with equal frequency at different points in the region. If the electron energy is a minimum, i.e., the electron is located at the lowest level ( $n = 0$ ), the particle is most often encountered at the centre of the "well"; if the electron is located in the state corresponding to  $n = 1$ , it is never encountered at the centre of the allowed segment, etc. The  $\psi_n^2$  curves indicate the frequency with which the electron appears at various points in the region.

Now, let us summarise. On the basis of quantum mechanics, the following conclusions may be drawn regarding the motion of a microparticle in a square well. Only motion corresponding to a discrete set of energy values,  $\mathcal{E}_0, \mathcal{E}_1, \mathcal{E}_2 \dots$ , is possible. The particle cannot be stationary since even the lowest energy level corresponds to motion with a certain velocity. Data on the nature of particle mo-

tion for a given energy are provided by the  $\psi^2(x)$ . Knowing  $\psi^2(x)$ , one may determine at what points in the region the microparticle appears often, and at what points rarely.

It remains to be determined under what conditions an "ordinary", i.e. classical, description of particle behaviour is valid.

Imagine that an oxygen molecule is enclosed in a box the dimensions of which are scores of times greater than the dimensions of the molecule. Assume that the molecule has the average energy, at room temperature, of a molecule of oxygen gas, i.e.,  $10^{-13}$  erg. Using the values  $a = 100 \text{ \AA}$ ,  $m = 5.4 \times 10^{-23} \text{ g}$  and  $\mathcal{E} = 10^{-13} \text{ erg}$ , one may determine the quantum level of the microparticle. The result is  $n = 1,000$ . Two conclusions may be drawn. First, the  $\psi_n^2$  curve has such a large number of alternating maxima and minima that the probability density of the particle is approximately the same for all points in the box. Secondly, neighbouring energy levels are very close.

Both characteristics which follow from the fundamental equation of quantum mechanics have disappeared: the probability distribution of the particle is practically indistinguishable from a smooth curve and the energy levels are so close that energy discreteness is imperceptible. In such a case, quantum mechanics yields approximately the same results as particle mechanics. This is true whenever the particle energy corresponds to a large quantum number. We have illustrated an important principle of quantum mechanics: when the quantum number is large, the results of quantum mechanics coincide with those of the mechanics of "ordinary" particles. This means that when  $n$  is large the particle-path concept and other characteristics of ordinary particles are applicable to microparticles as well.

#### Sec. 187. SIGNIFICANCE OF THE SOLUTION OF THE SCHRÖDINGER EQUATION

We have devoted a relatively large amount of space to the motion of a particle in a square well. On the basis of this simple example, we were able to illustrate the basic features of the quantum-mechanical method of considering problems. If an electron or other particle is able to execute motion in a limited region, the characteristic features of the solution of the Schrödinger equation are preserved, no matter what the form of the potential curve in this region. In all cases, the potential well may be intersected by a number of horizontal lines—possible energy levels. In principle, the Schrödinger equation enables us to calculate these energy values if the form of the potential well is given. The lowest level gives the zero-point energy of a particle in a given potential well.

For each energy level of number  $n$ , quantum mechanics establishes a set of wave functions  $\psi_n(x, y, z)$ . The quantity  $\psi_n^2(x, y, z)$  gives the probability of finding a particle at a given point in the region if the energy of the particle is  $\mathcal{E}_n$ . Since the particle manages to be at all points in the region more than once during the time of measurement,  $\psi^2(x, y, z)$  may be viewed as the density of the "particle cloud". The electron cloud surrounding an atomic nucleus resembles a photograph of the atom taken with long time-exposure. The  $\psi$ -function gives the amplitude of the wave associated with the particle. In the case of an electron in a square well, the  $\psi$ -function consists of standing waves and to every level there corresponds a characteristic wavelength  $\lambda$ .

This is not the situation in the general case. The standing "waves" corresponding to a given state (given  $n$ ) will appear very strange: their wavelengths  $\lambda = \frac{h}{\sqrt{2m(\mathcal{E} - U)}}$  will differ at different points in the region, depending on the nature of the potential "curve"  $U(x, y, z)$ . For more or less complex cases, there

is only slight similarity between the  $\psi$ -function and the amplitude of a standing wave (in the usual sense of the term).

Theoretical and experimental evidence indicate that in a number of cases several  $\psi_n$ -functions may correspond to a single  $\mathcal{E}_n$  energy value. This occurs if at one energy a particle may have states which differ with respect to another physical quantity, e.g., angular momentum. The forms of the  $\psi^2$  clouds of such a state—called a *degenerate* state—may differ radically from one another.

The problem of particle motion in a given type of potential well is solved when the energy levels are found and the characteristic  $\psi$ -functions are calculated for all levels. If the solution of the Schrödinger equation is known, the result of one or another measurement performed on the given particle may be predicted.

#### Sec. 188. TUNNELLING THROUGH A BARRIER

We shall now discuss a peculiar effect which may occur in the case of a microparticle, but not in the case of an ordinary particle. This is the tunnel effect, i.e., the “leakage” of a particle through a potential barrier.

Imagine that inside the region in which a particle moves, there is a potential barrier of height  $U$  and width  $d$  (Fig. 212). If the energy of the particle is  $\mathcal{E} < U$ , an ordinary particle could be either in front of the barrier or beyond the barrier. The particle cannot pass through the barrier, since this would require that the particle have a negative kinetic energy and an imaginary velocity, which is absurd. Matters stand differently with respect to microparticles. The uncertainty principle does not permit us simultaneously to ascribe to a microparticle exact values of velocity and coordinate, and hence of kinetic and potential energy. That is why a particle having a total energy  $\mathcal{E}$  may pass through the barrier.

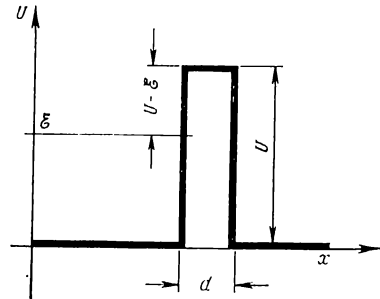


Fig. 212

The conditions for such passage may be determined as follows. Coordinate and momentum uncertainties are connected by the relation  $\Delta x \Delta p \approx h$ . Momentum uncertainty is uniquely related to kinetic energy uncertainty since  $K = \frac{p^2}{2m}$ . If  $\Delta K$  is a quantity of the order of  $U - \mathcal{E}$ , where  $\mathcal{E}$  is the particle energy and  $U$  is the height of the barrier, a particle to the left of the barrier (see the figure) has a coordinate uncertainty

$$\Delta x \sim \frac{h}{\Delta p} = \frac{h}{\sqrt{2m(U - \mathcal{E})}}.$$

If the barrier width  $d$  is less than  $\Delta x$ , the particle may be located on the other side of the barrier. The particle tunnels its way, so to speak, through the barrier at the total energy level  $\mathcal{E}$ .

Thus, the condition for tunnelling is

$$d \sqrt{2m(U - \mathcal{E})} < h, \quad \text{i.e.,} \quad \frac{d}{h} \sqrt{2m(U - \mathcal{E})} < 1.$$

It may be easily shown by numerical examples that the phenomenon is of significance only for microparticles.

For  $U - \mathcal{E} = 10 \text{ eV} \sim 10^{-11} \text{ erg}$ ,  $m \sim 10^{-27} \text{ g}$  (electron mass) and  $d \sim 10^{-8} \text{ cm}$ ,  
 $\frac{d}{h} \sqrt{2m(U - \mathcal{E})} = 0.2 < 1$ , i.e., tunnelling is possible.

For a 1-g spherule lying next to a match box ( $U - \mathcal{E} = 3,000 \text{ ergs}$  and  $d = 2 \text{ cm}$ ),  
 $\frac{d}{h} \sqrt{2m(U - \mathcal{E})} = 2.5 \times 10^{28} > 1$ . It is evident that the spherule cannot "tunnel" through the match box.

The probability of leakage through a barrier may be calculated. It turns out that this probability is proportional to

$$e^{-\frac{4\pi}{h} \sqrt{2m(U - \mathcal{E})} d}.$$

The tunnel effect could be predicted on the basis of the Schrödinger equation. The solution of this equation shows that even at points where  $U > \mathcal{E}$  the  $\psi$ -function has values differing from zero. Thus, there is a certain probability—inversely proportional to the magnitude of  $U - \mathcal{E}$ —that an electron is located in a region where in the language of "ordinary" particles it possesses negative kinetic energy.

## Atomic Structure

## Sec. 189. ENERGY LEVELS OF A HYDROGEN ATOM

A hydrogen atom has one electron which "rotates" in a nuclear field. One would think that the problem is simple. But even for this most simple atom, the solution of the Schrödinger equation is very cumbersome and, therefore, will not be reproduced here. However, we shall give the results of the calculations and discuss them in considerable detail.

An electric force of Coulomb attraction acts between the electron and the nucleus. The potential energy of an electron in a nuclear field is  $U = -\frac{e^2}{r}$ , where  $e$  is the charge of an electron (the same as the charge of a proton) and  $r = \sqrt{x^2 + y^2 + z^2}$  is the distance between the electron and the nucleus. The Schrödinger equation has the form

$$\Delta\psi + \frac{8\pi^2m}{h^2} \left( \mathcal{E} + \frac{e^2}{r} \right) \psi = 0.$$

Such an atom constitutes a peculiar kind of potential well and is illustrated in Fig. 213. This is a well without a bottom and with divergent sides. The sides of the well are hyperbolas and the axis  $r = 0$  is one of the asymptotes. The electron inside the atom has a negative potential energy\* since the minimum value of potential energy tends to infinity when  $r \rightarrow 0$  and the maximum value is equal to zero.

Fig. 214 shows the energy levels obtained from the solution of the Schrödinger equation. An important feature of the solution is the drawing together of the levels as the quantum number  $n$  increases. Transitions between levels will be discussed below. The scales of values, which are proportional to energy, are given in the units adopted in spectroscopy: volts and reciprocal centimetres. The energy level formula may be written in the form

$$\mathcal{E}_n = -\frac{2\pi^2me^4}{h^2n^2}.$$

For historical reasons, it is customary to write this formula in the form

$$\mathcal{E}_n = -\frac{cRh}{n^2},$$

where  $R = \frac{2\pi^2me^4}{ch^3} = 109,740 \text{ cm}^{-1}$  is the *Rydberg constant*.

Thus, it transpires that the total energy as well as the potential energy of the atomic electron is negative. The atomic electron may be located at any one of  $n$  levels. The greater the value of  $n$ , the higher the energy level, and the greater the energy possessed by the electron. The electron of a free hydrogen atom on which no force acts is at the lowest energy level:  $\mathcal{E}_1 = -cRh$ .

---

\* It may be asked: why has the origin of potential energy been chosen in such a manner that the electron energy is negative? The advantage of such a choice is not difficult to see. For different atoms, the potential energy has the same value only when  $r \rightarrow \infty$ . It is natural to set this common value equal to zero.

If an energy greater in magnitude than  $cRh$  is transmitted to the atom the electron leaves the bounds of the potential well, i.e., the atom becomes ionised. The energy  $cRh$  is called *the ionisation energy*.

It is customary to characterise the work of tearing away an electron from an atom by *the ionisation potential*. For a hydrogen atom,

$$V_{ion} = \frac{cRh}{e} = \frac{3 \times 10^{10} \times 109.740 \times 6.6 \times 10^{-27}}{4.8 \times 10^{-10}} = 4.5 \times 10^{-2} \text{ CGS units} = 13.5 \text{ V.}$$

The reason for the above designation is the following. Let us assume that the electron is torn away from the hydrogen atom by the action of a beam of electrons.

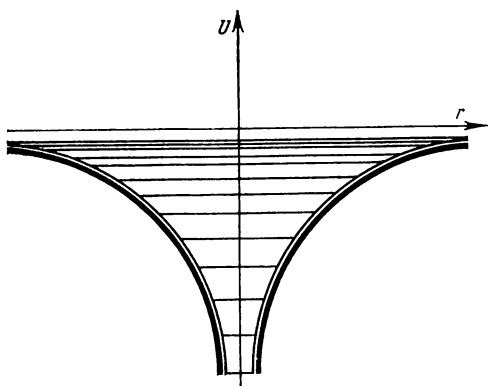


Fig. 213

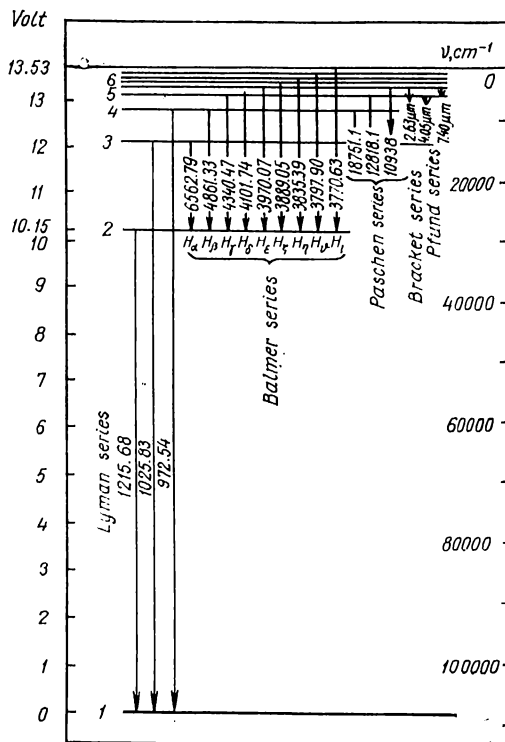


Fig. 214

To ionise a hydrogen atom, one must accelerate electrons, which act as projectiles, to an energy of at least  $eV = cRh$ . Therefore,  $V$  is the potential difference through which an electron must be accelerated in order to produce ionisation of collision with a hydrogen atom.

If the energy imparted to a hydrogen atom is less than  $cRh$ , a transition of the atom occurs to one of the  $n$  levels\*. Such an atom is said to be in an excited state.

An atom stays in an excited state for a small fraction of a second and then passes to a lower level with the emission of a photon in accordance with the equation

$$h\nu_{mn} = \mathcal{E}_m - \mathcal{E}_n = cRh \left( \frac{1}{n^2} - \frac{1}{m^2} \right).$$

\* In the case of a hydrogen atom, which has one electron, the phrases "the atom is at an energy level  $n$ " and "the electron is at an energy level  $n$ " have the same meaning.

If hydrogen atoms are excited by different kinds of collisions, they are raised to different energy levels and return to the ground state by “skipping” over levels (see Fig. 214). Therefore, a large concentration of hydrogen atoms will radiate photons of every possible  $\nu_{mn}$  frequency. A characteristic line spectrum of emission arises.

By calculating for a given  $n$  the  $\nu_m$  frequencies corresponding to the numbers  $m = n + 1, n + 2, \dots$ , we obtain a series of frequencies of lines in the hydrogen

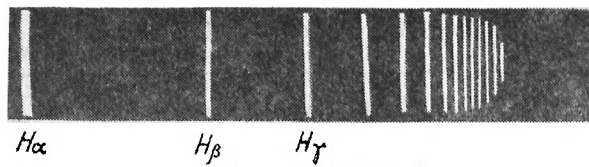


Fig. 215

spectrum. The existence of such series was known long before quantum mechanics was developed. The series corresponding to  $n = 2$  is known as the Balmer series. By substituting  $m = 3, 4, 5, 6, 7$  and  $8$ , respectively, in the formula, let us calculate the wavelengths of six of the lines in this series:  $\lambda_3 = 6,562.80 \text{ \AA}$ ;  $\lambda_4 = 4,861.38 \text{ \AA}$ ;  $\lambda_5 = 4,340.51 \text{ \AA}$ ;  $\lambda_6 = 4,101.78 \text{ \AA}$ ;  $\lambda_7 = 3,970.11 \text{ \AA}$ ; and  $\lambda_8 = 3,889.09 \text{ \AA}$ . It is apparent that the separation between lines decreases as  $m$  increases, which is in accordance with experimental results (see Fig. 215). Experimental and calculated values do not differ by more than  $0.05 \text{ \AA}$ .

#### Sec. 190. QUANTUM NUMBERS

The solution of the Schrödinger equation enables us to determine all of a hydrogen atom's energy levels,  $\mathcal{E}_n$ , as well as all of its wave functions. In the ground state, an electron is characterised only by the function  $\psi_1$ . As for the excited states, they are degenerate to the square of  $n$ , to use the terminology of quantum mechanics. This means that there are four  $\psi$ -functions corresponding to the energy  $\mathcal{E}_2$ , nine corresponding to  $\mathcal{E}_3$ , etc. Each of these states may actually exist.

How do the  $n^2$  states having the same quantum number  $n$  differ from one another? Quantum mechanics provides the answer to this question. States with one and the same energy value  $\mathcal{E}_n$  may differ with respect to the magnitude of the electron's angular momentum as well as the value of the angular momentum's projection on a certain selected axis.

The solution of the Schrödinger equation for a hydrogen atom shows that the electron's angular momentum has a discrete series of values given by the formula

$$L = \sqrt{l(l+1)} \frac{h}{2\pi},$$

where  $l$  may assume any integral value from  $0$  to  $n - 1$  when the electron is at the  $n$ -th level.

Moreover, the Schrödinger equation shows that relative to the selected direction  $z$  the angular momentum  $L$  must be oriented in such a manner that

$$L_z = m \frac{h}{2\pi},$$

where  $m$  is an integer that may assume any value from  $-l$  to  $+l$ , including zero.

It should be recalled that according to the uncertainty principle  $L$  and  $L_z$  give us all that we can possibly know about the angular momentum; in other words, it is meaningful to specify simultaneously only these two quantities.

Thus, the state of an electron in an atom is characterised by three quantum numbers:  $n$ ,  $l$  and  $m$ . The number  $n$  is called the *principal* quantum number,  $l$  the *azimuthal* quantum number, and  $m$  the *magnetic* quantum number.

The states with  $l = 0, 1, 2, 3, \dots$  are designated by the letters  $s, p, d, f, \dots$ , respectively. The principal quantum number precedes one of the above letters. For example, the  $3p$  state is the state with  $n = 3$  and  $l = 1$ .

Let us list all of the possible states for  $n = 1, 2$  and  $3$ :

$n$	$l$	Designation of the state	$m$
1	0	$1s$	0
2	0	$2s$	0
	1	$2p$	-1, 0, 1
3	0	$3s$	0
	1	$3p$	-1, 0, 1
	2	$3d$	-2, -1, 0, 1, 2

The energy transitions of a hydrogen atom are determined exclusively by the values of the principal quantum number  $n$ . In order for the  $l$  and  $m$  numbers to play a part, the degeneracy must be "removed", i.e., the energy of states with a different angular momentum must be changed. In the case of hydrogen atoms,

this may be done by placing the atoms in a magnetic field. In other cases, degeneracy is removed by electron interaction (see below).

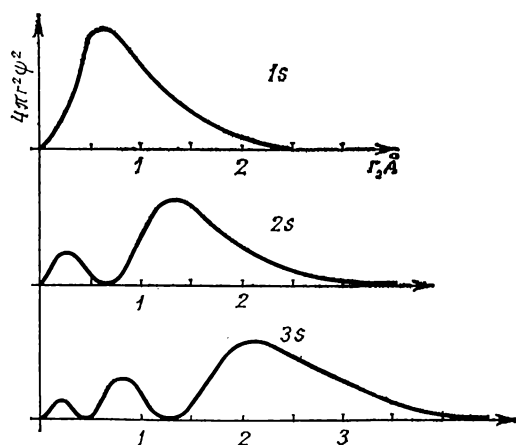


Fig. 216

there is only one  $\psi$ -function. The equation  $l = 0$  means that the electron has no angular momentum. This requires, of course, that there be no favoured directions of motion, i.e., the electron cloud must have spherical symmetry. Such is the result obtained from the Schrödinger equation: the functions  $\psi_{1s}, \psi_{2s}, \psi_{3s}$ , etc., have spherical symmetry.

Fig. 216 shows curves of radial density distribution of the electron cloud or, what amounts to the same, the probability density distribution of the electron. The quantity  $4\pi r^2 \psi^2$ , which gives the radial density, is plotted along the ordinate.

#### Sec. 191. THE ELECTRON CLOUD OF $s$ AND $p$ STATES

The state characterised by the three numbers  $n, l$  and  $m$  is described by the wave function  $\psi_{n, l, m}$ . The characteristic form of the electron cloud corresponding to this state is determined by the function  $\psi_{n, l, m}^2$ . Let us consider the form of the  $\psi^2$ -functions of a hydrogen atom which characterise its various excitation states.

Consider the  $s$  states. Since  $l = 0$ ,  $m$  is also equal to zero. Hence, for every  $n$



It is evident that  $4\pi r^2 \psi^2 dr$  represents the number of electrons\* contained in a spherical shell the inner and outer radii of which are  $r$  and  $r + dr$ , respectively. The radial density curves show that in the  $1s$  state there is one electron density maximum, which for a hydrogen atom is located at a distance of  $0.53 \text{ \AA}$  from the nucleus. In the  $2s$  state, there are two density maxima; the electron will be within the second maximum most of the time. Finally, in the  $3s$  state, there are three density maxima; here, the electron will be within the third maximum most of the time.

As the principal quantum number  $n$  increases, the electron cloud becomes dissipated.

The  $p$ -state functions look entirely different. There are three values of  $m$ , namely,  $0$ ,  $-1$  and  $+1$ , corresponding to  $l = 1$ . The electron-cloud configurations are illustrated in Fig.

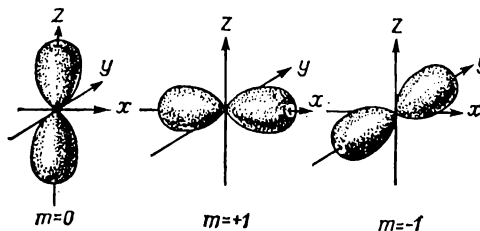


Fig. 217

217. For  $m = 0$ , the major axis of the "figure eight" is oriented along the selected direction; for  $m = \pm 1$ , the major axis is perpendicular to the selected direction. It is evident that the  $m = \pm 1$  states may be meaningfully distinguished only when both are present. The figure gives some indication of the symmetry of the electron cloud. It is the same for all states. A change in the principal quantum number merely results in a change in the nature of the radial drop in density: the greater the value of  $n$ , the more extended the curve.

We shall not discuss states with large values of  $l$ . Their electron clouds are more complex.

## Sec. 192. PAULI'S EXCLUSION PRINCIPLE

Atoms are arranged in the Mendeleyev periodic table in accordance with the number of electrons contained in them. Thus, helium has two electrons, lithium three and beryllium four. On the basis of the Schrödinger equation, what can be said about the structure of atoms?

At first glance the problem may appear hopeless. Even in the case of helium, strict adherence to procedure would involve solving the Schrödinger equation for a wave function with six variables,  $\psi(x_1, y_1, z_1, x_2, y_2, z_2)$ , the square of which gives us the probability of finding the first electron at point  $x_1, y_1, z_1$  when the second electron is at point  $x_2, y_2, z_2$ . The potential energy to be substituted in the equation is

$$U = -\frac{2e^2}{r_1} - \frac{2e^2}{r_2} + \frac{e^2}{r_{12}},$$

where  $r_1$  and  $r_2$  are the distances of the electrons from the nucleus (the nuclear charge of helium is  $2e$ ) and  $r_{12}$  is the distance between electrons. An exact solution of such a problem is quite impossible.

It would be extremely desirable to deal separately with each atomic electron and describe each such electron by its wave function  $\psi(x, y, z)$ . But how can this be done? Evidently, it is necessary to consider the motion of one electron in the

\* A fractional number of electrons should not disturb us, for this is only a manner of speech. Strictly speaking,  $4\pi r^2 \psi^2 dr$  is the probability of an electron being inside a spherical shell of  $dr$  thickness.

field of the nucleus and the remaining electrons. It may be assumed that this effective field has spherical symmetry. Therefore, the description of the properties of such an electron will not differ from that of the electron of a hydrogen atom.

To be sure, the problem is still quite difficult: for different electrons these effective fields differ and, moreover, all must be determined simultaneously since each depends on the states of the remaining electrons. Such an effective field is said to be self-consistent. This approach to the problem of a multi-electron atom enables us to apply, to a large extent, the description of the properties of a hydrogen atom's electron to the behaviour of an electron of a complex atom.

The state of each electron will be characterised by the same quantum numbers as in the case of hydrogen. However, in the case of a multi-electron atom, degeneracy is removed by electron interaction, and levels with different  $l$  and  $m$  values will have different energies.

The Schrödinger equation enables us to determine the energy levels that are possible, but does not indicate the energy of the atomic electrons. One might think that all the electrons of an atom occupy the lowest energy level. In any case, such would be the behaviour of "ordinary" particles. But experiments completely refute such a supposition. The "arrangement" of electrons in accordance with energy levels is governed by the *Pauli exclusion principle*. The first conjecture that such a principle exists was based on a study of the Mendeleyev periodic table.

As was indicated above, it follows from the Schrödinger equation that  $(2l + 1)$  states exist for a given  $n$  and  $l$ . It was also indicated that this in turn yields  $n^2$  different  $\psi$ -functions for a single value of  $n$ . The first values of  $n^2$  are 1, 4 and 9. Let us examine the Mendeleyev periodic table. Helium, neon and argon, which complete the first three periods of the table, contain 2, 8 and 18, i.e.,  $2n^2$ , electrons, respectively. This is by no means accidental. It is an expression of a profound law according to which only two electrons can have the same  $\psi$ -function. In other words, an energy level cannot be occupied by more than two electrons. This general law of nature, to which we shall return in Sec. 194, is called the Pauli exclusion principle.

By means of this principle, we can "arrange" the electrons of a complex atom in accordance with quantum numbers and, therefore, in accordance with energy levels and values of angular momentum. A helium atom has two electrons, which may occupy the single  $1s$  level. The third electron of lithium must be located at the next level, i.e., the  $2s$  level. A beryllium atom has four electrons, which occupy the  $1s$  and  $2s$  levels. The fifth electron of boron is at the  $2p$  level. At this level, there are six places for electrons, which are all filled when neon is reached. But let us postpone consideration of the relation of the Mendeleyev periodic law to the electron structure of an atom until Sec. 196.

Do two electrons occupying a level which is characterised by the same three quantum numbers differ in any way? It turns out that two such electrons differ with respect to the orientation of their internal angular momentum ("spin" orientation).

#### Sec. 193. DEFLECTION OF AN ATOMIC BEAM IN A MAGNETIC FIELD

In the preceding articles, the angular momentum of an electron due to its motion about a nucleus was discussed in considerable detail. The presence of such angular momentum may be demonstrated since an atom acquires a magnetic moment as a result of the motion of its electron.

Let us employ classical concepts and assume that the electron revolves in a circle of radius  $r$ . Since current is equal to the charge transferred in a unit time,

the equivalent current of such a revolving electron is  $I = ne$ , where  $n$  is the number of revolutions per second. On the other hand,  $n = \frac{v}{2\pi r}$ , where  $v$  is the velocity. Thus,

$$I = \frac{ve}{2\pi r},$$

and the magnetic moment of the revolving electron (see p. 198) is

$$M = \frac{1}{c} IS = \frac{1}{c} \frac{ve}{2\pi r} \times \pi r^2 = \frac{1}{2c} evr.$$

$L$ , the angular momentum of an electron, is equal to  $mvr$ . Hence,

$$M = \frac{e}{2mc} L.$$

This relationship between the angular momentum and the magnetic moment of an electron moving about a nucleus, obtained by means of the above simple calculations, has been experimentally confirmed for all atomic electrons.

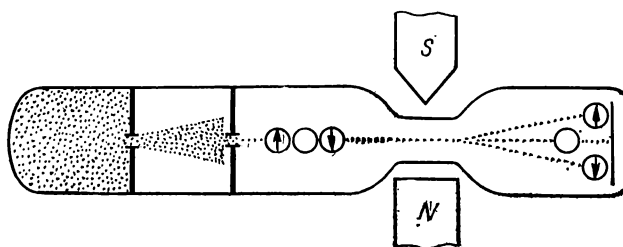


Fig. 218

Thus, atoms for which  $L \neq 0$  possess magnetic moments, and it may be demonstrated in a suitable experiment that such atoms behave like small magnets. Fig. 218 illustrates such an experiment in which a parallel beam of atoms passes through a nonuniform magnetic field.

Air is carefully pumped out of a long tube. At the left end, an atomic gas is created. Atoms in the left compartment can get out through small apertures in the screens. Since there are two apertures, only atoms moving along the axis of the apparatus will remain in the beam. This parallel beam of atoms passes through a nonuniform magnetic field. As was indicated on p. 217, in such a field the force acting on a body which has a magnetic moment  $M$  is

$$f = M_z \frac{dH}{dy},$$

where  $\frac{dH}{dy}$  is the field gradient in the direction perpendicular to the atomic beam and  $M_z$  is the projection of the magnetic moment on the direction line of the gradient. If the magnetic moment is perpendicular to the field, no force acts on the body, but if the moment is directed along the field, the body is attracted to either the north or south pole, depending on the direction in which  $M$  is pointed. If the atomic beam contains atoms having different magnetic moments, or differently oriented magnetic moments, the beam of atoms will spread out: atoms travel in different directions and different forces  $f$  act on them. The beam is allowed to impinge on a plate until enough atoms hit the plate to be perceptible.

The above experiment has been of great importance in the development of basic atomic theory. It is still of importance as a method of determining the magnetic moments of atomic nuclei.

#### Sec. 194. ELECTRON SPIN

By means of experiments with atomic beams, one can measure  $M$  and, therefore,  $L$ . One of the important deductions of quantum mechanics pertains to the quantisation of angular momentum:  $L$ : the total momentum  $L$  can assume only a discrete set of values, namely,

$$L = \sqrt{l(l+1)} \frac{h}{2\pi},$$

where  $l$  is the azimuthal quantum number. Therefore, the magnetic moments of atoms also can assume only a discrete set of values:

$$M = \frac{eh}{4\pi mc} \sqrt{l(l+1)} = \mu \sqrt{l(l+1)}.$$

The coefficient in this formula,

$$\mu = \frac{eh}{4\pi mc} = 0.927 \times 10^{-20} \text{ erg/Gs},$$

is called the *Bohr magneton*.

In experiments with atomic beams, the external magnetic field is directed perpendicular to the atomic beam. The projections of  $L$  on this perpendicular line can also have only a discrete set of values:

$$L_z = m \frac{h}{2\pi}$$

hence,

$$M_z = m\mu,$$

where  $m$  is the magnetic quantum number. It is seen that the values of  $M_z$  must be equal to a whole number of Bohr magnetons.

$M_z$  may be determined directly from experiments with atomic beams. What should be the result of experiments with beams of atoms? We may expect the following picture. Hydrogen, helium, lithium and beryllium have only  $s$ -electrons. Since  $L = 0$  for these atoms, such a beam does not split up in a magnetic field. When  $p$ -electrons are present, the beam may be expected to split up into three distinct components: an undeviated central component for  $m = 0$ , and two components arranged symmetrically to the right and left for  $m = \pm 1$ . For  $d$ -electrons, we should obtain five beam components corresponding to the five possible values of the quantum number  $m$ , etc.

These predictions are realised to the following extent: in certain cases a beam of atoms does not split up, while in others it splits up into distinct components. Thus, it is evident that some atoms have no magnetic moment; if an atom has a magnetic moment, it is quantised.

As regards the results obtained for specific atoms, they point to a completely new fact, namely, that an electron has an internal magnetic moment.

Such is the conclusion to be drawn from an experiment with a beam of hydrogen atoms: a beam of hydrogen atoms splits up into two symmetrical components, which correspond to the deflections of atoms with magnetic moments  $\pm\mu$ . There is no undeviated central component. This fact may be explained by the following

hypothesis, which is supported by other data: an electron has a magnetic moment and there are only two possible orientations of this moment in space, namely, the projections  $\pm\mu$  on the direction line of the external field.

The magnetic moment of an electron due to its motion about a nucleus is uniquely related, as we have just seen, to the angular momentum of the electron motion about the nucleus. It also transpires that the internal magnetic moment of an electron is related to its internal angular momentum, or *spin*.

As far back as 1925, before the above experiments which show that an electron has an internal magnetic moment were performed, Goudsmit and Uhlenbeck suggested that an electron has a spin. These physicists showed that such a hypothesis—the existence of an electron spin, i.e., an internal angular momentum—removes insurmountable difficulties in deciphering spectra. At first, it was proposed that spin is a consequence of the rotation of an electron about its own axis (whence the origin of the term “spin”). But this interpretation is incorrect. Electron spin is a primary characteristic and is not reducible to something simpler.

What is the relationship between the internal angular momentum (spin) and the internal magnetic moment of an electron? The experiment with a beam of hydrogen atoms leads to the conclusion that  $M_z$ , the projection of the internal magnetic moment of an electron, may assume only two values, namely,  $\pm\mu$ . It may be assumed that  $L_z$ , the projection of the spin, may also assume only two values.

If the formula

$$L = \sqrt{l(l+1)} \frac{h}{2\pi}$$

is applied to the internal angular momentum of an electron, one finds that the number  $l$  has a single value. Thus, as quantum mechanics indicates, the values  $l = 0, 1, 2, \dots$  correspond to 1, 3, 5, ..., and in general  $(2l+1)$  states, respectively. To obtain two spin states, which the experimental results show to be the case ( $2l+1 = 2$ ), one must assume that  $l = \frac{1}{2}$ .

The absolute value of an electron's internal angular momentum (spin) has the single possible value  $L = \sqrt{\frac{3}{4}} \frac{h}{2\pi}$ . As for the projection of spin, assuming as before that the differences in the possible values of  $L_z$  must be multiples of  $\frac{h}{2\pi}$ , we see that it can assume only two values, namely,  $+\frac{1}{2} \frac{h}{2\pi}$  and  $-\frac{1}{2} \frac{h}{2\pi}$ . Thus,  $(L_z)_{sp} = s \times \frac{h}{2\pi}$ , where  $s$  is a new quantum number (*spin number*) that can assume only two values, namely,  $\pm \frac{1}{2}$ .

It was stated earlier that the hydrogen experiment led to deflections corresponding to a magnetic moment of one magneton, and that  $M_z = \mu$ . Since the quantum number  $s$  is equal to  $\frac{1}{2}$ , the relation

$$M = \frac{e}{2mc} L$$

turns out to be incorrect for the internal motion of an electron. Agreement with experimental results is obtained if

$$M_{sp} = \frac{e}{mc} L_{sp}.$$

Thus,  $\frac{M}{L}$ , the ratio of magnetic moment to angular momentum for electron motion about a nucleus, is one half of the analogous ratio for internal motion of an electron.

The existence of electron spin enables us to formulate the Pauli exclusion principle more clearly. There can be no more than two electrons in a state having quantum numbers  $n$ ,  $l$  and  $m$ . Such electrons differ only with respect to spin projection. Experiments show that the spin projections of two such electrons cannot be the same. The Pauli exclusion principle may now be formulated as follows: there can be only one electron in a state characterised by *four* quantum numbers  $n$ ,  $l$ ,  $m$  and  $s$ . In other words, if there are two electrons in an  $n$ ,  $l$ ,  $m$  state, their spin directions are opposite to each other.

#### Sec. 195. MAGNETIC MOMENTS OF ATOMS

The electron spin hypothesis makes it possible to interpret the results of experiments with atomic beams. The measurement of magnetic moment is one of the most important methods of determining the electron state of atoms. Let us consider the first few elements of the Mendeleyev periodic table.

We have already studied the hydrogen atom. What happens to a beam of helium atoms? Such a beam does not split up. This is as it should be. This atom has two electrons in the  $2s$  state. The Pauli exclusion principle requires that their spins, and hence their magnetic moments, be oppositely directed; the total magnetic moment is equal to zero.

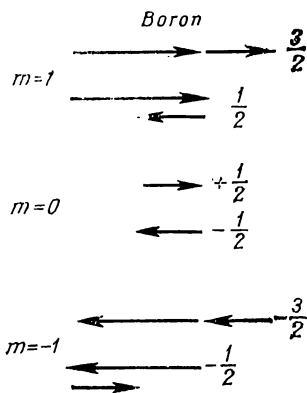


Fig. 219

In a lithium atom, the magnetic moment must be determined by the third electron since the actions of the spins of the first two electrons in the  $1s$  state annul each other. The third lithium electron is in the  $2s$  state. Therefore, like in the case of hydrogen, the magnetic moment can be determined only by the spin. The splitting up of a lithium beam into two components shows that this analysis is correct. Like in the case of hydrogen, one component corresponds to a spin projection  $+\frac{1}{2}$ , and the other to a spin projection  $-\frac{1}{2}$ .

A beryllium atom has four electrons—two in the  $1s$  state and two in the  $2s$  state. Since the spins of the electrons in each pair are oppositely directed, they annul each other and yield a total moment of zero. Therefore, a beam of beryllium atoms does not split up into components.

However, a beam of boron atoms splits up into four components. How can this be explained? It is evident that the magnetic moment is produced only by the fifth electron, which is in the  $2p$  state. This state may be realised with three values of the magnetic quantum number, namely,  $+1$ ,  $0$  and  $-1$ . Thus, the orbital magnetic moment may have three values, including zero. The values of the spin magnetic moment must be added to those of the orbital magnetic moment. How is this done? Fig. 219 shows all the possible mutual orientations of the moments. It is seen that four combinations of angular momentum are possible:  $\frac{3}{2}$ ,  $\frac{1}{2}$ ,

$-\frac{1}{2}$  and  $-\frac{3}{2}$  (in  $\frac{h}{2\pi}$  units, usually not indicated).

The splitting up of a beam of carbon atoms is even more complex: seven lines, including one which is not deflected, with the following values of angular momentum:  $3$ ,  $2$ ,  $1$ ,  $0$ ,  $-1$ ,  $-2$  and  $-3$ .

## Sec. 196. THE MENDELEYEV PERIODIC LAW

The regularities of the Mendeleyev periodic system become understandable when viewed in the light of available data on the electron structure of atoms.

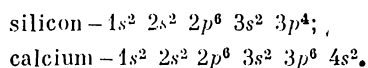
The basic rules of distribution of electrons in an atom according to quantum numbers are derived from the Pauli exclusion principle. But if we were governed by this information alone, only the electrons of the first eighteen atoms (through argon) would be assigned correct quantum numbers. This may be explained as follows. Electrons fill the energy states of an atom in consecutive order. In a hydrogen atom, the energy levels are degenerate, the energy being determined by only one quantum number. In multi-electron atoms, degeneracy is removed as a result of interaction and the energy of an electron in an atom depends on all the quantum numbers. It goes without saying that in a hydrogen atom an electron at a level for which  $n = 3$  has less energy than an electron at a level for which  $n = 4$ . This is not necessarily so in multi-electron atoms when the azimuthal quantum number is large. Thus, in a number of cases, the "natural" consecutive order of quantum numbers does not correspond to the order in which the energy states of an atom are filled.

It is customary to group the electrons of an atom into shells, whereby electrons having the same principal quantum number are said to belong to the same shell. The shells are usually designated by the following letters:

$$K, L, M, N, \dots,$$

which correspond to  $n = 1, 2, 3, 4, \dots$ , respectively.

The electron cloud of an atom is described in the main by the distribution of electrons according to quantum numbers or shells. This distribution is represented by formulas which indicate by a superscript the number of electrons having the same  $n$  and  $l$ . For example:



To determine to what extent a shell is complete, one should remember that the maximum number of electrons in the subshells  $s, p, d, f, \dots$  is 2, 6, 10, 14, ..., respectively. These values are obtained from the formula  $2(2l + 1)$ .

Returning to the Mendeleyev table, let us see where the order of distribution of electrons according to quantum numbers is violated. The first such violation occurs for potassium. The last electron is in the  $4s$  level rather than in the  $3d$  level. Calcium, the next element in the table, receives another  $4s$  electron. Then, beginning with scandium, the 21st element, the  $3d$  level is built up. But when we reach chromium, the 24th element in the table, a new anomaly arises. The order of distribution of quantum levels according to energies has changed. It becomes unfavourable from the energy viewpoint to have two electrons in the  $4s$  level. The configuration of chromium is therefore  $3s^2 3p^6 3d^5 4s$ .

We shall not discuss the remaining anomalies. The electron configurations for all elements may be found in any physics or chemistry handbook. The main point is the following: the distributions of electrons according to quantum numbers, which are explained by purely physical methods of investigation (spectral analysis and the measurement of magnetic moments), aid us in understanding the chemical properties of the various elements.

## Sec. 197. IONISATION POTENTIALS

One of the methods used to study the electron configuration of an atom is to measure its ionisation energy, i.e., the energy that must be expended in removing an electron from the atom. Since the energy of an electron in an atom is negative and is reckoned from zero (the energy of an electron removed from an atom), the ionisation energy is simply equal to the energy level occupied by the electron in the atom. It is customary to refer this energy to the electron charge and express it in volts. For example, we say that the ionisation potential of a hydrogen atom is equal to 13.53 volts. This means that to free an electron one must perform work equal to that of moving an electron through a potential difference of 13.53 volts. Fig. 214 shows the significance of this value.

In the case of a multi-electron atom, one may find a series of ionisation potentials which characterise the levels of the first, second, third, etc., electrons, calculating from the position of the least bound electron. In this sense, one speaks of the first, second, etc., ionisation potential of a given atom.

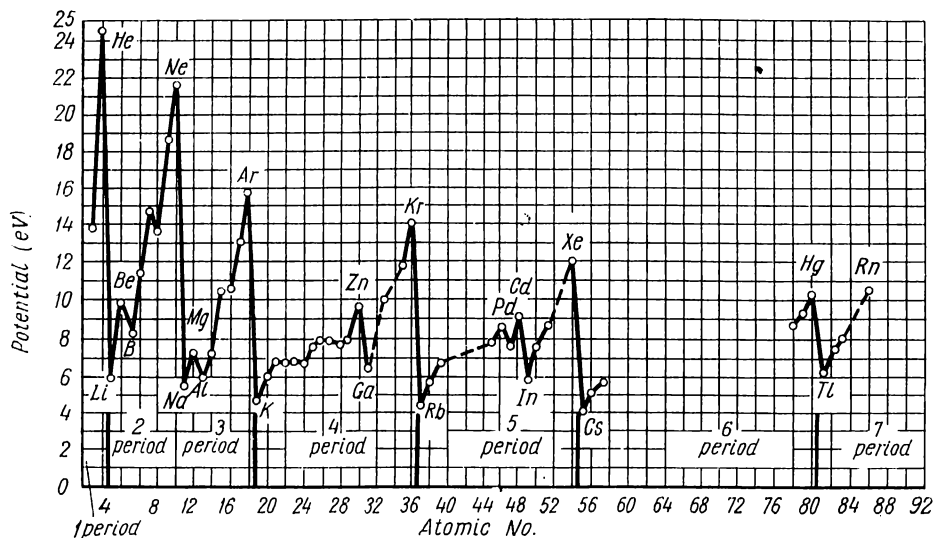


Fig. 220

There exist many methods of measuring ionisation potentials. For this purpose, gases or vapours are placed in an electric field. A stream of electrons emitted by a heated filament ionises the gas. As long as the energy of a primary electron is insufficient to dislodge an atomic electron, the electric current passing through the gas does not change. When the energy of the primary electrons becomes sufficiently great to dislodge electrons from the atomic gas, a considerable number of positively charged ions will be present in the region and a sharp increase in the electric current occurs. By gradually increasing the voltage applied to the apparatus, one can determine very accurately the instant when this increase in electric current begins. This critical value of voltage gives the magnitude of the ionisation potential. The values of the first ionisation potentials of most chemical elements are graphically illustrated in Fig. 220.

It is easily seen that the periodicity of this property completely conforms with that of the periodic table. It is most difficult to dislodge an electron from a helium



atom and atoms of the other noble gases. This is precisely the reason for their chemical inertness. The univalent alkali metals have the lowest potentials. This, too, is completely understandable to the chemist, who is familiar with the exceptional ability of such substances to enter into reactions.

The values of the first and successive ionisation potentials are related to the valency of the atoms. Atoms of the alkali elements are univalent because one electron, which is on the outer ring of the atoms of these substances, is more weakly bound than the rest of the electrons. The first potentials of a caesium atom, for example, have the following values: 3.9, 27, 46 and 62 volts. It is seen that the differences between the energies necessary to dislodge the first and succeeding electrons are quite large.

#### Sec. 198. ATOMIC SPECTRA IN THE OPTICAL REGION

Atomic absorption spectra as well as emission spectra may be obtained, but only the latter are of basic importance. Atomic spectra of emission in the optical region may be obtained by spectroscopic investigation of the radiation produced by gases and the vapours of bodies which are solid at normal temperatures.

In order for atoms to radiate, they must be excited, i.e., made to pass from a lower energy level to a higher one. When atoms return to lower energy levels, an emission spectrum is produced. For every transition, there is a corresponding line in the spectrum.

Atoms may be excited by various means. One method consists in the use of a gas discharge. The voltage applied to a gas-discharge tube accelerates the charged particles in the gas. These particles collide with neutral atoms, to which energy is transmitted by impact. Another method, which is used in the spectral analysis of metals, consists in the creation of arcs or sparks between two electrodes made of the material under investigation. Very high temperatures are produced in an arc or spark, resulting in the vaporisation of the substance in the region of the discharge. The atoms are excited as the result of collisions.

An atomic emission spectrum consists of a very large number of sharp lines. The radiation frequency corresponding to a given line satisfies the equation  $h\nu_{mn} = E_m - E_n$ . Thus, by measuring the frequencies of radiated light, we can determine the differences in the energy levels of a given atom. One can reliably interpret atomic spectra, i.e., determine energy level patterns, from the values of radiation frequencies. Handbooks provide data on the spectral lines and energy levels of the chemical elements.

It should not be supposed that a spectrum contains lines corresponding to all transitions from any one level to any other. Experiments have confirmed, and a theoretical basis has been provided for, the fact that certain selection rules exist. Certain transitions are forbidden, i.e., they do not exist.

One cannot predict, of course, to which lower energy state an excited atom will pass, and what will be the frequency of the radiated spectral line. But not all transitions occur with equal probability. In principle, the probability of transition from one level to another may be theoretically calculated. The magnitude of this probability determines, in the main, the intensity of the corresponding spectral line.

Atomic spectra are affected by external fields. If the substance under investigation is located in an electric or magnetic field, a number of its spectral lines split up into several components. The energy of a system having a magnetic moment  $M$  and located in an external magnetic field  $H$  is given by the expression  $U = -MH$  (see p. 208). States which have the same quantum numbers  $n$  and  $l$  may

differ from each other with respect to the projection of the magnetic moment on the direction line of the magnetic field. Therefore, the application of a magnetic field removes the degeneracy of energy levels and atomic electrons having different magnetic quantum numbers will have different energies.

Investigations of atomic spectra of emission in the optical region are of great practical importance. Such a method of spectral analysis constitutes a very sensitive means of determining the chemical composition (up to  $10^{-10}$  g) of substances, primarily alloys, and in a number of cases is more sensitive than chemical analysis.

Optical frequencies usually arise with relatively weak excitation of an atom, i.e., when outer, valence electrons are transferred to a higher level. But even a very "high" electron can produce a broad spectrum. It would appear that the radiation frequency has no lower limit. Thus, the energy level diagram shows that as  $n$  increases the levels come closer together (Fig. 214 shows the levels and transitions for hydrogen, but in principle the patterns are the same for other atoms). This means that transitions corresponding to very low frequencies (long wavelengths) occur. However, experiments show that spectra produced by outer electrons, even though they extend into the infrared region, do not include lines of very long wavelength. It must be concluded that the probability of a transition to some energy level such as the 21st is not large, and the probability of a transition from the 21st to the 20th, for which a photon of low  $\nu$  would be radiated, is quite negligible.

In the direction of high frequencies (short wavelengths), the frequency is limited by the ionisation potential. With respect to the "highest" electron, the potential of helium is the greatest and that of caesium the lowest, viz., 24 V and 4 V, respectively. This corresponds to radiation frequencies of  $6 \times 10^{15}$  Hz ( $\lambda = 500 \text{ \AA}$ ) and  $10^{15}$  Hz ( $\lambda = 3,000 \text{ \AA}$ ), respectively. Thus, only a high-level electron can bring us into the region of very short ultraviolet wavelengths, which, relative to characteristic X-ray radiation, may also be called a region of very long wavelengths.

It is quite understandable that electrons of inner shells can be raised to high levels with strong excitation. In such a case, the characteristic spectrum includes X-rays.

#### Sec. 199. ATOMIC X-RAY SPECTRA

In multi-electron atoms, the ionisation potentials of low levels reach high values. The excitation of such atoms may, therefore, result in the radiation of X-rays (wavelengths of the order of  $0.1\text{--}10 \text{ \AA}$ .) An energy of the order of  $10^4$  eV must be imparted to an atom in order to produce X-ray radiation. This may be achieved in gas-discharge tubes by applying a voltage of tens of thousands of volts.

One may calculate the value of the temperature at which an atom begins to radiate X-rays due to thermal collisions with other atoms. If the average kinetic energy per degree of freedom is to be of the order of  $10^4$  eV, the temperature must be of the order of  $10^8$  K. Such high temperatures are achieved in solar and celestial atomic explosions (see p. 449). The X-ray radiation of the Sun may be determined by means of instruments placed in artificial Earth satellites.

A practical method of obtaining X-rays is by the bombardment of a solid (the anti-cathode of an X-ray tube) with a stream of electrons. Electrons impinging on the anti-cathode are abruptly braked. As a result, a continuous spectrum of X-rays is obtained. The electron energy, which has been increased to a value  $\mathcal{E}_1$  by acceleration in an electric field, decreases to a value  $\mathcal{E}_2$  as the result of braking. The energy difference  $\mathcal{E}_1 - \mathcal{E}_2 = h\nu$  is released in the form of radiation.

$\mathcal{E}_2$  may assume any value from  $\mathcal{E}_1$  to zero. Hence the radiation frequencies lie between the limits  $\nu = \frac{\mathcal{E}_1}{h}$  and zero. The electron energy which does not go into radiation is transformed into heat. Only about one-hundredth of the energy of the electron beam is transformed into X-ray energy. Evidently, the continuous X-ray spectrum has a short-wavelength limit  $\lambda_{\min} = \frac{c}{\nu_{\max}} = \frac{hc}{eV}$ . Substituting the values of the constants, we obtain

$$\lambda_{\min} = \frac{12.3}{V},$$

where  $\lambda$  is expressed in angstroms and  $V$  in kilovolts. Beginning at a very definite wavelength, the continuous X-ray spectrum increases in intensity with increasing wavelength, reaches a maximum several score angstroms from the short-wavelength limit, and then slowly decreases in intensity.

Investigations show that sharp lines, having a characteristic form for every element, are superimposed on the continuous spectrum. A characteristic X-ray

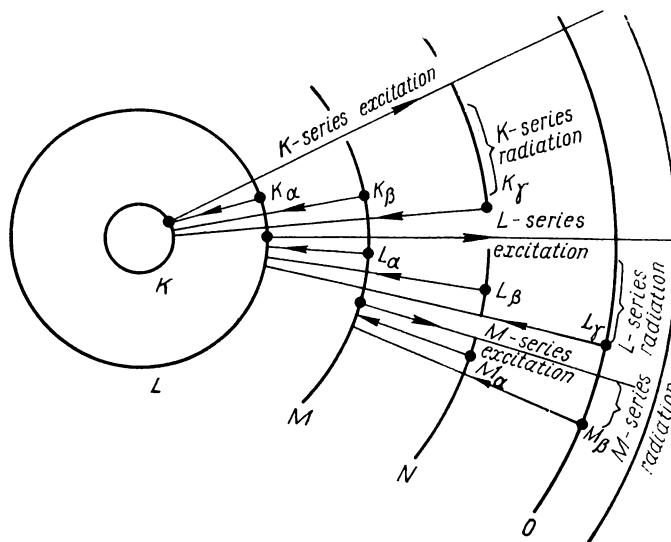


Fig. 221

spectrum arises owing to the fact that some of the electrons which impinge on the anti-cathode penetrate the atoms and dislodge inner electrons, i.e., electrons in the  $K$ ,  $L$ , etc., shells. An X-ray quantum is produced when a high-level electron passes over to a vacated low-level position. The set of spectral lines due to electron transitions to the  $K$  level is called the  $K$  series, to the  $L$  level the  $L$  series, etc. If the voltage applied to an X-ray tube is increased, the series will appear in consecutive order because, as the energy of the electrons impinging on the anti-cathode is increased, more and more low-energy levels will be consecutively vacated and made available for transitions. The  $K$  series will be the last to appear.

The general scheme of electron X-ray transitions is shown in Fig. 221, where heavy dots indicate initial levels. The most intense lines are marked on the diagram. However, some transitions are missing since they are forbidden by the selection rules. This pertains, for example, to transitions with the same value of azimuthal quantum number.

Since the configuration of completed lower shells is the same for all atoms, X-ray spectra diagrams of different atoms are very similar to each other. All spectra contain typical sequences of lines, which are systematically shifted along the wavelength scale in accordance with the atomic number of the elements. For example, all elements produce a strong  $\alpha$  doublet ( $K_{\alpha_1}$  and  $K_{\alpha_2}$ ) and a weaker  $\beta$  doublet. Quite often these doublets are unresolved. In such a case, one speaks of the  $\alpha$  line and  $\beta$  line of a given element's  $K$  series. These doublets are of a "spin" nature.

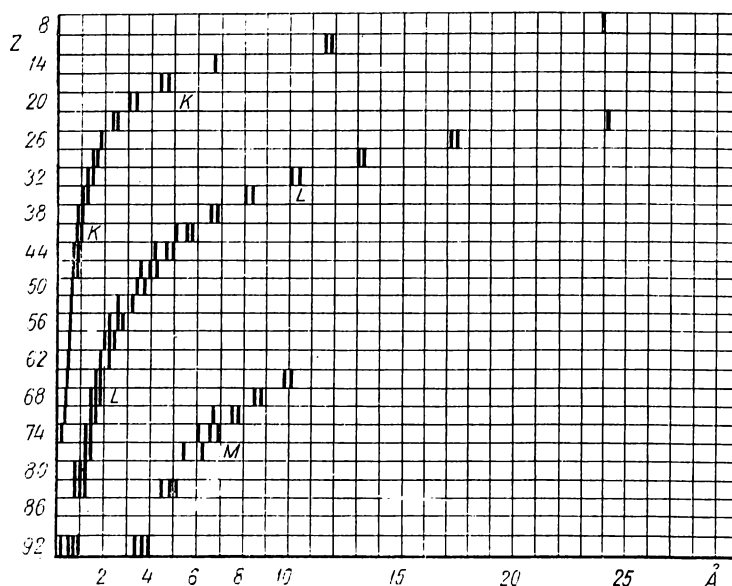


Fig. 222

Fig. 222 shows that a characteristic spectrum is gradually shifted into the short-wavelength region with increasing atomic number of the element giving rise to this spectrum. This law was discovered by Moseley. Its physical basis is the steady increase of the interaction force between an electron and a nucleus with increasing nuclear charge. The formula expressing this law will not be presented, but the systematic displacement of the lines is clearly evident from the figure.

# Molecules

## Sec. 200. CHEMICAL BONDS

A molecule is a stable configuration of atoms. Every atom in a molecule occupies a stable position. The displacement of an atom in any direction results in an increase in the potential energy of the molecule. When an atom approaches a neighbouring atom there is a force of repulsion, and when it recedes a force of attraction. Every atom of a molecule, and the molecule as a whole, is in a potential well.

The form of the potential curve of an atom or molecule is quite evident (see Fig. 223). Since it is not possible to reduce the distance between atoms to zero the curve of potential energy as a function of their separation rises sharply as this distance decreases. In the direction of increasing separation, the curve rises from the equilibrium position, i.e., the bottom of the well, much less sharply. Variations are possible: the potential energy at great distances may be more or less than at the bottom of the well and the well may have or may not have a clearly defined wall. The energy of a molecule may be more or less than the sum of the energies of its atoms. Accordingly, when atoms are combined in a molecule, heat is either released or absorbed (see p. 435).

An atom in a potential well is bound to its neighbours. What is the reason for this bond? Do various types of bonds exist? *Ionic* and *homopolar* bonds are two ideal classifications of chemical bonds. In the overwhelming majority of cases of interest in chemistry, one of these two types of bonds, or an intermediate case in which both ideal types coexist, occurs.

If an atom can transfer one or several electrons to another atom electrostatic attraction will occur between the ions which are formed. This is what is meant by an ionic bond. At a certain interatomic separation which is characteristic of this pair of ions, the forces of electrostatic attraction are counterbalanced by the repulsion of the electron clouds of the atoms.

If an atom is to transfer an electron to another atom, it is necessary that this process be advantageous from the energy viewpoint. In such a case, the simple tendency to pass over to the lowest energy level will result in the transfer of an electron.

It was shown on p. 380 that an electron can be torn away from a neutral atom by an expenditure of energy equal to the product of an electron charge and the ionisation potential of the atom. Thus, the formation of a positive ion is always associated with an expenditure of work. On the other hand, the formation of a negative ion, i.e., the attachment of an electron to a neutral atom, is associated with a release of energy. To be sure, this applies only to the first electron. The attachment of a second electron to a singly charged negative atomic ion requires an expenditure of work to overcome the electrostatic repulsion.

An ionic bond can exist if the energy of tearing away an electron, i.e., the work of creating a positive ion, is less than the sum of the energy released in the forma-

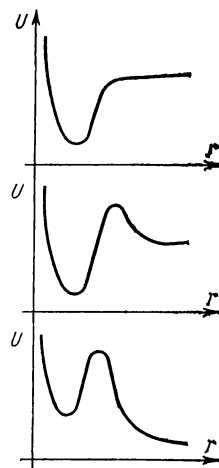


Fig. 223

tion of negative ions and the energy due to electrostatic attraction between the ions.

The alkali metals, in which the last electron is just beginning to form a new shell, have the lowest ionisation potentials. In the alkaline earth metals, each atom has two loosely bound electrons. It is evident that the formation of a positive ion from a neutral atom requires the least work when the electrons to be torn away are just beginning to form a new shell.

On the other hand, it turns out that the most energy is released when an electron becomes attached to a halogen atom, in which the outer shell is one electron short of being complete. Therefore, in a large number of cases, an ionic bond is formed when a transfer of electrons resulting in the creation (in the formed ions) of closed electron shells, characteristic of atoms of the noble gases, occurs. In this way, the physical significance of potential wells in such molecules as NaCl and  $\text{MgCl}_2$  may be easily explained.

However, this explanation is not valid in all cases. For example, diatomic molecules of hydrogen, oxygen, etc., are not covered by this explanation. It cannot be assumed that in uniting one of these atoms is transformed into a negative ion and the other into a positive ion. Theoretical arguments need not be mustered. The physical properties of molecules formed of ions indicate whether an ionic bond exists or not. Specifically, ionic compounds dissociate and form electrolytes. A large class of organic molecules do not behave in this manner. Therefore, for such substances, an ionic model is clearly not applicable.

How can one explain the bond between atoms of such molecules? We must determine whether or not a gain in energy occurs when, say, two hydrogen atoms unite to form a molecule.

Such a gain does take place, and the conditions for its occurrence are given by quantum mechanics. As was stated on p. 379, the electron of a hydrogen atom behaves, in the main, like an electron in a potential well. The zero energy level of an electron in a potential well is determined by the dimensions of the well (see p. 374), i.e., the smaller its dimensions the greater the zero-point energy. Thus, any expansion of the region in which the electron could move results in a decrease in energy.

Now, imagine that two hydrogen atoms, which have one electron each, come into contact with each other. Since the Pauli exclusion principle allows two electrons to be in one state, the regions in which the electrons exist may merge and create a potential well of increased dimensions. This can occur only for two electrons having opposite spins.

If a third atom approaches the hydrogen molecule which has formed, the argumentation used above is no longer applicable. The third electron cannot merge its region of motion with that of the electrons in the hydrogen molecule since this is not allowed by the Pauli exclusion principle: the vacant sites of the hydrogen molecule are occupied by two electrons of opposite spin.

Thus, the second type of bond, a so-called homopolar bond, is provided by a pair of electrons of opposite spin. In the case of an ionic bond there is a transfer of electrons from one atom to another, but here the bond is achieved by joint action of the electrons, i.e., it is as if a common region of motion has been created. An expansion of the region in which an electron may move results in a decrease in energy and this is the reason for the formation of a potential well. This bond—the merging of the electronic clouds of electrons having opposite spins—is the main type of bond in organic molecules.

Each atom is capable of forming a limited number of homopolar bonds. Two electrons of opposite spin, having a common “living space” in the form of the overlapping clouds of their wave functions, take part in the creation of each bond.

As we know, *s* electrons have spherically symmetrical  $\psi$ -functions, but the  $\psi$ -functions of *p*, *d* and *f* electrons extend in specific directions. Therefore, a homopolar bond between any two electrons, except *s* electrons, will be a directed bond. If a bond has formed between two atoms, the electronic clouds of these atoms assume a definite orientation relative to the first bond line. Thus, only certain specific angles are formed between bond lines emanating from these atoms. The values of such normal bond angles may be derived from quantum mechanics for all atoms.

To a certain extent, both types of bonds are ideal. We frequently encounter cases in which the physical and chemical properties of a molecule make it necessary to adopt an intermediate bond mechanism. In an ionic bond an electron is completely transferred from one atom to another and in a homopolar bond each electron belongs equally to both bound atoms, but in intermediate cases the electrons implementing a bond may spend more time near one of the atoms than the other. Such a model reflects, for example, the existence of an ionic bond in which the bond electrons belong most of the time to a negative ion and the existence of a homopolar bond in which the bond electrons spend almost the same amount of time with each of the bound atoms. Intermediate bonds of any percentage of "ionocity" are possible.

### Sec. 201. GEOMETRIES OF MOLECULES

A vast amount of data on spacings between the centres of atoms in molecules and crystals has been accumulated. Most of this data has been obtained by diffraction methods. If we do not insist on a very high degree of accuracy, it turns out that it is possible to represent molecules by models which give the shape and dimensions of the molecules.

Models of molecules of the NaCl type, in which the atoms are joined by an ionic bond, are particularly simple. Each ion may be represented by a sphere having a definite radius. The dimensions of a number of ions are given in the following table:

Ion	Li <sup>+</sup>	Na <sup>+</sup>	K <sup>+</sup>	Cs <sup>+</sup>	F <sup>-</sup>	Cl <sup>-</sup>	Br <sup>-</sup>	I <sup>-</sup>
Ion radius, Å	0.60	0.95	1.33	1.69	1.36	1.81	1.95	2.16

By means of such a table, we can determine the spacing between the centres of ions in any salt. For example, in NaCl it is equal to  $0.95 + 1.81 = 2.76$  Å.

But what is the significance of the assertion that an ion may be represented by a sphere? To show that such a representation is justified, we must determine how closely to a molecule (say NaCl) another ion (sodium or chlorine) may approach. This is possible since experiments indicate that both fused and solid salts consist of ions. It turns out that a second and a third ion approach a given ion just as closely as the first. Moreover, ions the charges of which have the same sign may also approach each other to a distance equal to the sum of the ion radii. Thus, ions behave like spheres.

An important conclusion regarding ionic molecules may be drawn from these geometric facts. Let us assume that a group of molecules are gathered closely around one of the molecules. The arrangement of ions is shown in Fig. 224. Since complete equality exists in interatomic spacings, it is no longer possible to say with which chlorine neighbour of a given sodium ion, or with which sodium neigh-

hour of a given chlorine ion, a molecule is formed. The concept of a molecule has lost its meaning.

We must conclude from the geometric arrangement of atom centres that in a concentrated state, i.e., in a liquid or solid, where the atoms are linked by ionic bonds, molecules do not exist as distinct formations. The concept of a molecule turns out to be inapplicable.

But what is the situation in the case of a gas? Upon vaporisation, a pair of ions of opposite charge, the net electric charge of which is equal to zero, is most easily torn away from a liquid. Therefore, basically, molecules of the NaCl type are to be found in vapours. However, in addition to molecules, ions are also vaporised from a liquid.

The situation is entirely different in the case of molecules having homopolar, i.e., covalent, bonds. An analysis of interatomic spacings encountered in molecules shows that spacings between atom centres may be calculated by means of so-called atomic radii. The values of such radii in angstroms for the most often encountered atoms are as follows:

C—	C=	C≡	H—	O—	O=	O≡
0.771	0.665	0.602	0.30	0.66	0.55	0.50

Atomic radii decrease with increasing multiplicity of the valent bond. The table shows that the separation between two bound carbon atoms, C—C, is 1.54 Å, the separation in C—H is 1.07 Å, etc.

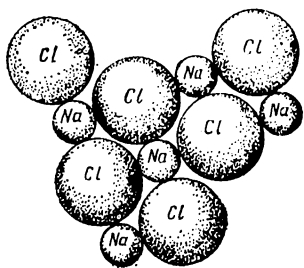


Fig. 224

In constructing a model of a molecule, we also have at our disposal certain elementary data on bond angles. The existence of normal bond angles is understandable from considerations of symmetry and is in agreement with certain qualitative quantum mechanical reasoning discussed in the preceding article. Thus, the normal bond angle of a carbon atom that is linked to four atoms is a tetrahedral angle ( $109^{\circ}28'$ ). In the case of an aromatic carbon atom, as well as other carbon atoms that are linked to three atoms, the normal bond angle is equal to  $120^{\circ}$ . Finally, the characteristic bond angle of a carbon atom that is linked to two atoms is  $180^{\circ}$ .

The normal bond angles of oxygen, sulphur and nitrogen atoms which are linked to two atoms in the case of oxygen and sulphur and to three atoms in the case of nitrogen are equal to  $90^{\circ}$ . The nitrogen atom in the nitro group  $\text{NO}_2$  has a normal bond angle of  $120^{\circ}$ .

In a number of cases, bond angles deviate considerably from the "normal". In certain cyclic compounds of the cyclobutane type, the angles are equal to  $90^{\circ}$  rather than  $109^{\circ}28'$ . Such deviations are due to spatial obstacles. However, before discussing this, we must clarify a third geometric characteristic of a molecule, namely, its intermolecular radius.

Investigations of molecular arrangements in crystals have shown that each atom may be assigned an intermolecular radius, such that on the average neighbouring molecules will touch each other. Thus, for example, the intermolecular radius of hydrogen is 1.17 Å, oxygen—1.36 Å, nitrogen—1.57 Å, etc. This does not mean, however, that the distances between atoms of the same molecule which are not linked by valent bonds are determined by these values. The dimensions and the form of a molecule are determined by interaction between the forces establishing



equilibrium distances between atoms which are not linked by valent bonds and the forces establishing normal bond angles. Since the bond forces between atoms are an order of magnitude greater than the other forces, the interatomic distances do not change and the configuration of the molecule is determined by competition between the elasticity of a bond angle and the compressibility of the intermolecular sphere of an atom.

Here is a simple, but graphic example. Experiments show that the bond angle in a molecule of water is equal to  $105^\circ$ . The distance between hydrogen atoms is  $1.54 \text{ \AA}$ . Therefore, considerable compression of the intermolecular spheres of the

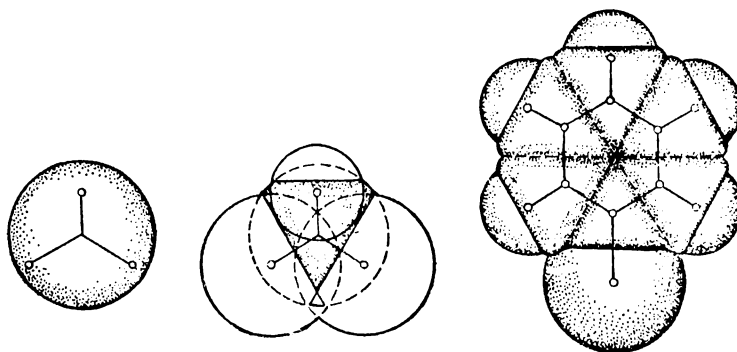


Fig. 225

hydrogen atoms occurs. This compression ( $2 \times 1.17 - 1.54 = 0.8 \text{ \AA}$ ) is balanced by the elasticity of the bond angle, the normal value of which is equal to  $90^\circ$ . Thus, the forces which compress the hydrogen atoms by  $0.8 \text{ \AA}$  are equal to the forces which change the angle from  $90^\circ$  to  $105^\circ$ . Such a simple mechanism explains the difference between the structure of a hydrogen sulphide molecule and that of a water molecule. Since the length of the hydrogen-sulphur bond is considerably longer than that of the hydrogen-oxygen bond, the hydrogen atoms of the former molecule are considerably less "crowded". It turns out that in hydrogen sulphide the distance between hydrogen atoms is equal to  $1.99 \text{ \AA}$  and the bond angle is equal to  $92^\circ$ . The compression of hydrogen atoms by  $0.35 \text{ \AA}$  is balanced by a change of only  $2^\circ$  in the bond angle. Many organic molecules may be used to illustrate the validity of this mechanism.

Fig. 225 shows the structure of a chlorobenzene molecule. To the left is a carbon atom; in the centre, the beginning of the build up, viz., a  $\begin{array}{c} \text{H} \\ | \\ \text{C} \\ / \quad \backslash \\ \text{C} \quad \text{C} \end{array}$  group, and to the right, a model of the molecule.

## Sec. 202. THE ELECTRONIC CLOUD OF A MOLECULE

Electron motion in a molecule, just as in an atom, is described by a wave function. Strictly speaking, the  $\psi$ -function is a function of  $3n$  coordinates, where  $n$  is the number of electrons in the molecule. Then,  $\psi^2$  will give the probability of any electron distribution, i.e., the "electron density".

It has been noted already that the solution of the Schrödinger equation for multi-electron atoms is very complex. In the case of molecules, the difficulties are, of course, even greater. Only approximate, semiempirical methods of calculation are applicable here. In this connection, physical methods of determining electron density are of particular importance. However, even when such methods are used, the results are quite limited.

The time-averaged electron density of a molecule is determined by means of X-ray structural analysis (the electron density gives the probability that the electrons are at a given location). As a result of the vibrations of atoms inside a molecule, and of the molecule as a whole, a photograph of an electronic cloud is smeared. Figure 167 (p. 296) shows a cross-section of the electron density pattern of an anthracene molecule. The coarseness of the method may be gauged from the fact that the hydrogen atoms of the molecule are not apparent in all cases. The method used in plotting the pattern is similar to that used in the construction of topographical maps. Electron peaks and valleys are indicated by lines connecting points with the same electron density. Each atom is represented by an electron density "hill". Superposition of the bell-shaped density functions of two atoms along a bond line results in the formation of a "bridge" between the atoms. Unfortunately, the accuracy of the method is too poor to enable us to determine the nature of the chemical bond by measuring the height of this bridge. Its height is indistinguishable from the sum of the density functions of two free atoms. But the specific nature of the chemical bond should probably be manifested in an additional increase in electron density (as compared with free electrons). Such electronic cloud patterns are, therefore, merely interesting illustrations of molecular structure.

If the electron density with respect to the atomic nuclei of a molecule were known, we would be able to calculate the dipole moment of the molecule. For this purpose, it would be necessary to determine the centres of "gravity" of the positive and negative charges. The dipole moment has not yet been determined in this manner, although comparison of neutronographic (neutrons scattered on nuclei) and roentgenographic data could be used to solve such a problem. However, the dipole moment of a molecule can be reliably measured (see p. 519) and it is then possible to solve the converse problem, namely, determine the centre of "gravity" of negative charge by means of the dipole moment.

It would seem that in purely ionic molecules we encounter the extreme case in which the centre of gravity of an electronic cloud coincides with the centre of an anion. The dipole moment of KCl, for example, could then be predicted in the following manner. If one electron is taken from a potassium atom and transferred to a chlorine atom, one "extra" positive charge will be separated from one "extra" negative charge by the distance between the potassium and chlorine centres, i.e.,  $1.81 + 1.33 = 3.14 \text{ \AA}$ . Hence, the dipole moment will be equal to  $3.14 \times 4.8 \times 10^{-18} = 15 \text{ CGS units}$ . But experiments yield a value of 6.8 CGS units. This means that even in the case of such a classical ionic bond the potassium electron does not go over completely to the anion. On the other hand, the other extreme case is fully realised. Evidently, symmetrical molecules such as  $\text{H}_2$ ,  $\text{O}_2$  and benzene cannot have a dipole moment: the centres of gravity of the electronic cloud and the nuclei coincide.

One other property of an electronic cloud should be mentioned, namely, its ability to be displaced relative to a nucleus. Electronic clouds may be displaced relative to nuclei by means of an electric field. Since nuclei are much heavier than electrons, it may be assumed that the nuclei remain fixed. The displacement of the electronic cloud of a molecule may be described by the displacement of its centre of gravity. When the centre of gravity of negative charge is displaced rela-

tive to the centre of gravity of positive charge by a distance  $x$ , the molecule acquires an induced dipole moment  $p = Nex$ , where  $N$  is the number of electrons in the molecule. The induced dipole moment increases linearly with the field, i.e.,  $p = \beta E$ . It is customary to describe the displacement of the centre of gravity of an electronic cloud by  $\beta$ , the magnitude of the polarisability of the molecule. The quantity  $\beta$  has the dimensions of volume. The greater the volume of the molecule, the greater the value of  $\beta$  (see Chapter 35).

### Sec. 203. ENERGY LEVELS OF MOLECULES

The energy of an atom changes only by one means: a change occurs in its electron motion, i.e., an electron passes into another quantum state. The energy of a molecule may also change in this manner, but by other means as well. For example, the atoms of a molecule vibrate relative to one another. The vibrational energy is an integral part of the energy of a molecule and also may assume only a discrete set of values. Furthermore, a molecule rotates as a whole. The rotational energy is also quantised and a change in the state of a molecule may result in a change in rotational energy. Therefore, the energy state of a molecule is described by indicating the state of its electronic cloud (electron level), the state of its vibrational motion (vibrational level) and the state of its rotational motion (rotational level). We deal with three kinds of information—analogueous, so to speak, to a house number, floor number and apartment number.

But which is analogous to a floor number and which to an apartment number? Which energy levels are separated by large intervals and which by small ones? The answers to these questions are contained in the energy level diagram shown in Fig. 226, which is based on experimental results and theory. Two electronic levels,  $e'$  and  $e''$ , are shown in this figure. Associated with each electronic level is a group of vibrational levels designated by a set of  $v$  values, and associated with each vibrational level is a group of rotational levels designated by a set of  $j$  values.

Clearly, the intervals between rotational levels are less than between vibrational levels, and those between vibrational levels are less than between electronic levels.

Let us assume that a molecule may have electronic levels at 100, 200, 300, ... energy units, vibrational levels at 10, 20, 30, ... units, and rotational levels at 1, 2, 3, ... units. In such a case, a molecule at the second electronic level, first vibrational level and third rotational level will have an energy of 213 units.

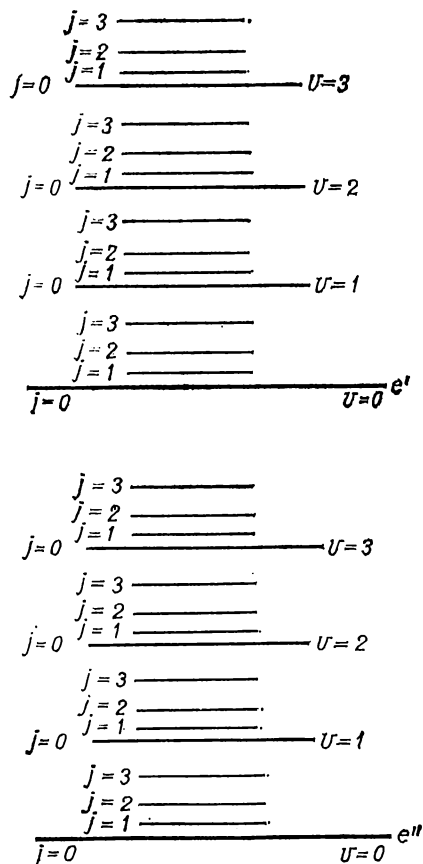


Fig. 226

Thus, the energy of a molecule may be given in the form\*

$$W = W_{el} + W_{vib} + W_{rot}.$$

The frequency of radiated or absorbed light may always be determined from the difference in energy between two levels, i.e.,

$$\nu = \frac{1}{h} (\Delta W_{el} + \Delta W_{vib} + \Delta W_{rot}).$$

It would be interesting to examine transitions involving a change in only one "kind" of energy. Practically, this is possible only for rotational transitions and it is easily seen why this is so.

Let us investigate the absorption of electromagnetic waves by a group of molecules. Beginning with the longest wavelengths, i.e., smallest packets of energy  $h\nu$ , we find that the molecules do not absorb energy as long as the magnitude of a quantum of energy is less than the difference between two neighbouring levels. By gradually increasing the frequency, we eventually obtain quanta of energy that are capable of raising molecules from one "rotational" level to another. Experiments show that this occurs in the microwave region (at the end of the radio band) or, in other words, in the far infrared spectrum. It is found that wavelengths of the order of 0.1-1 mm are absorbed by molecules. Thus, a pure rotational spectrum may be obtained.

By further increasing the frequency, we enable the rotational spectrum to become more developed, but nothing new occurs until the quanta of energy impinging on the substance are of sufficiently high frequency to make molecules pass from one vibrational level to another. It is clear, however, that a pure vibrational spectrum, i.e., a series of transitions for which the number of the rotational level does not change, is never obtained. Transitions from one vibrational level to another involve various rotational levels. For example, a transition from the zero (lowest) vibrational level to the first may be accomplished by molecules from the fourth rotational level to the third, the third to the second, etc. Thus, there arises a vibration-rotational spectrum, which may be observed in infrared light (3-50  $\mu\text{m}$ ). Clearly, all transitions from one vibrational level to another are close to one another and yield a group of very close lines in the spectrum. For low resolution, these lines merge into one band. Each band corresponds to a definite vibrational transition.

By increasing the frequency still further, we finally reach a new spectral region, which is characteristic of a molecule. This occurs in the optical and ultraviolet portion of the spectrum where the energy of a quantum suffices for the transition of a molecule from one electronic level to another. Here, of course, neither pure electronic transitions nor pure electronic-vibrational transitions are possible. Electronic-rotational transitions, involving a change in "house", "floor" and "apartment", occur. Since a vibration-rotational transition gives rise to a band, the spectrum in the optical region is "striped", i.e., it consists of a system of bands.

Now, let us discuss the various types of molecular spectra in detail.

#### Sec. 204. THE ROTATIONAL SPECTRUM OF MOLECULES

Free rotation of molecules occurs only in gaseous state. Therefore, basic data on rotational energy levels are obtained by studying gas spectra. Investigation of these spectra by optical means is very difficult. Much more suitable for this pur-

\* For one more summand of the energy of a molecule and energy absorption related with it see Sec. 215.

pose is a radio-spectroscopic procedure that has been developed during recent years. A generator of electromagnetic waves transmits radiation through a wave-guide\* which is partially filled with the gas under investigation. After passing through the gas, the electromagnetic waves arrive at a receiver which measures their intensity. This measurement may be performed over a large range of frequencies. The width of the band of frequencies generated by radio methods may be made so narrow that the resolving power becomes hundreds of thousands of times (!) greater than in the case of optical methods. Optical methods enable us to distinguish lines separated by  $0.1 \text{ cm}^{-1}$ , but by radio methods we can distinguish lines separated by  $10^{-6} \text{ cm}^{-1}$ \*\*. By means of this high resolving power, we are able to solve a number of interesting problems which are discussed below. A rotational spectrum arises as a result of the quantisation of a molecule's kinetic energy of rotation:

$$K_{\text{rot}} = \frac{I\omega^2}{2}$$

where  $I$  is the moment of inertia of the molecule. This is the form of the expression for the energy of a diatomic molecule. This energy is described by a single moment of inertia taken about an axis perpendicular to the line joining the atoms and which passes through the centre of inertia. As was indicated earlier, in the general case the rotation is described by three moments of inertia taken about three main axes.

Briefly, let us consider the rotational spectra of diatomic molecules.

First, it should be emphasised that not all molecules, including diatomic molecules, will yield a rotational spectrum of radiation or absorption. As has been explained already (see p. 243) every radiator or absorber of electromagnetic waves is a kind of oscillator, i.e., an elementary dipole. If the atomic motion of a molecule or the motion of a molecule as a whole is not accompanied by a change in dipole moment, such motion cannot result in the radiation or absorption of electromagnetic waves.

When a molecule radiates or absorbs energy, its dipole moment  $p$  varies periodically as the oscillation frequency. The dipole moment oscillates about an average value, corresponding to the equilibrium position of the atoms. It may be shown that the intensity of the spectral lines is proportional to the derivative  $\left(\frac{dp}{dr}\right)_{r=r_0}$ , i.e., the maximum rate of change of the dipole moment with respect to interatomic spacing. All symmetrical molecules the atoms of which are joined by homopolar bonds have a constant zero value of  $p$ . Therefore, they do not give rise to rotational spectra. Such molecules include, for example, all diatomic molecules of the same atoms ( $\text{H}_2$ ,  $\text{O}_2$ ,  $\text{N}_2$ , etc.).

Let us consider the rotational spectrum of a diatomic polar molecule, i.e., a molecule possessing a dipole moment. The rotational energy of such a molecule is  $K_{\text{rot}} = \frac{I\omega^2}{2}$ ; here  $\omega$  is the angular velocity of rotation and  $I$  the molecule's moment of inertia:

$$I = m_1 r_1^2 + m_2 r_2^2 = \frac{m_1 m_2}{m_1 + m_2} r^2,$$

where  $r_1$  and  $r_2$  are the distances to the centre of inertia and  $r = r_1 + r_2$ . The value of  $\omega$  is determined from the fact that according to a rule of quantum mechan-

\* A waveguide is a metallic duct of rectangular or circular cross-section through which centimetre radio waves may be propagated with practically no losses.

\*\* In spectroscopy, in addition to wavelength units, it is customary to use a reciprocal wavelength unit (wave number), i.e., the number of waves per centimetre.

ics (p. 382) the rotational momentum,  $I\omega$ , may assume only the discrete set of values

$$\frac{h}{2\pi} \sqrt{j(j+1)}$$

where  $j = 0, 1, 2, \dots$  is the quantum number designating the rotational levels. Therefore, the angular velocities of rotation of a molecule may assume only the following set of values:

$$\omega_j = \frac{h}{2\pi I} \sqrt{j(j+1)};$$

hence

$$K_{rot} = \frac{I\omega^2}{2} = \frac{h^2}{8\pi^2 I} [j(j+1)].$$

Beginning with a zero energy of rotation, the energy of successive levels increases in accordance with a square law.

Energy transitions are subject to a simple selection rule, i.e., only transitions between neighbouring levels are allowed (Fig. 227).

The radiation or absorption frequency in the rotational spectrum of a diatomic molecule is given by

$$\nu = \frac{hj}{4\pi^2 I} \quad (j = 0, 1, 2, \dots)$$

for a transition between the  $j$  and  $j - 1$  levels. In this simple case, the rotational spectrum consists of a system of equally spaced lines.

For different gas temperatures, the average energy of rotation of a molecule differs. In accordance with Boltzmann's law, the most probable energy is given

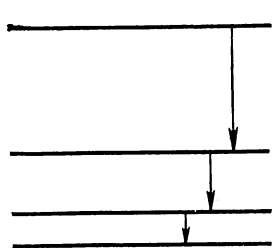


Fig. 227

by  $\frac{I\omega^2}{2} = kT$  (two rotational degrees of freedom—see p. 145). Thus, the number of the energy level at which a molecule is most frequently located may be easily calculated. For example, in the case of a molecule of a hydrochloric acid vapour ( $I = 2.61 \times 10^{-40}$  g cm<sup>2</sup>), at temperatures of 300, 600 and 1,200 K, we obtain  $j = 4, 6$  and 8, respectively.

Since transitions are possible only between neighbouring levels, a series of equally spaced frequencies will be grouped about the line of "average"  $j$ -value. The line intensity decreases as its distance from this  $j$ -value increases, since the number of molecules in the corresponding energy state decreases.

Rotational spectra enable us to determine interatomic distances in simple molecules to a very high degree of accuracy (much greater accuracy than by diffraction methods). Thus, if the number of atoms in a molecule is not large, the distances between atoms may be determined if the moment of inertia and the masses of the atoms are known. For a diatomic molecule,

$$r = \sqrt{\frac{I}{m}}, \quad \text{where} \quad m = \frac{m_1 m_2}{m_1 + m_2}.$$

In the case of a hydrochloric acid molecule:

$$m_H = 1.67 \times 10^{-24} \text{ g}, \quad m_{Cl} = 35 \times 1.67 \times 10^{-24} \text{ g}.$$

The separation between the H and Cl atoms in an HCl molecule is

$$r = \sqrt{\frac{2.61 \times 10^{-40} \times 36 \times 1.67 \times 10^{-24}}{35 \times 1.67 \times 10^{-24}}} = 1.63 \times 10^{-8} \text{ cm}.$$

This value agrees closely with values obtained by other means.

## Sec. 205. INFRARED VIBRATION-ROTATIONAL SPECTRA

This type of spectrum may be observed in a wavelength band extending from 2-3 to several score microns. For brevity, the vibration-rotational absorption spectrum is referred to as the "infrared spectrum". In the case of solids, where there is no molecular rotation, a pure vibrational spectrum is obtained. In the case of liquids, where rotation is impeded, the rotational structure of the band is smeared.

**Diatomic Molecules.** Let us disregard rotation for the present and consider vibrational energy levels.

The vibration of a diatomic molecule may be visualised by means of a simple model consisting of two spheres joined by a spring. In such a system, the natural frequency of oscillations is given by

$$\nu = \frac{1}{2\pi} \sqrt{\frac{k}{m}},$$

where  $k$  is the stiffness coefficient determining the binding force and  $m$  is the mass of an atom when the atoms in the molecule are the same; when the masses differ,  $m$  is the reduced mass, which is equal to  $\frac{m_1 m_2}{m_1 + m_2}$  (we leave the proof of this to the reader). Quantum mechanics shows that the energy of an oscillator is given by the formula

$$E = \left( \nu + \frac{1}{2} \right) h\nu.$$

Here,  $\frac{1}{2} h\nu$  is the zero-point energy of the oscillator, i.e., the oscillation energy at absolute zero, and  $\nu = 0, 1, 2, \dots$  is the oscillation quantum number. Moreover, it is shown in quantum mechanics that in the case of harmonic oscillators energy transitions may occur only between neighbouring levels. In the case of nonharmonic oscillators transitions skipping one level or more occur, but these are weaker than the main transitions. Harmonic oscillations occur under the action of a restoring force  $-kx$ . The potential energy of such oscillations is  $\frac{kx^2}{2}$ , i.e., the shape of the curve is parabolic.

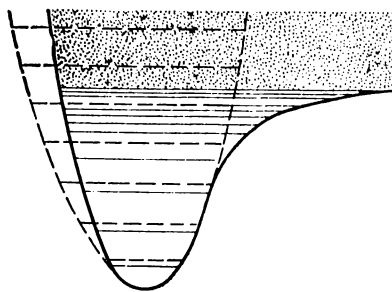


Fig. 228

Fig. 228 shows a potential curve (and an inscribed parabola) for a diatomic molecule. The horizontal lines represent energy levels based on theoretical calculations. For low values of energy, the deviation of the potential curve from a parabola is negligible. Such a molecule may be expected to obey the harmonic oscillator law as long as the vibrational energy is much less than the dissociation energy of the molecule. Under such conditions, the vibrational levels may be considered to be equally spaced, and since only transitions between neighbouring levels are allowed, the diatomic molecule will possess a single transition frequency. If there is no molecular rotation, the entire spectrum will consist of a single line. Actually, in addition to the main frequency  $\nu$ , the spectrum contains the "overtone" frequencies  $2\nu$ ,  $3\nu$ , etc. (as the separation between levels decreases, the proportional trend of the overtone frequencies is lost). However, the overtones are weak and in very many cases we have a right to speak of a single vibration frequency.

The presence of molecular rotation will transform such a spectral line into a band. If a molecule vibrates and rotates simultaneously, its energy is determined by the two quantum numbers  $\nu$  and  $j$ :

$$\mathcal{E} = \left( \nu + \frac{1}{2} \right) h\nu_{vib} + \frac{h^2}{8\pi^2 I} j(j+1).$$

The frequencies obtained now fall into two groups, one less than and one more than the vibration frequency  $\nu_{vib}$ . These groups are known as branches and are designated by the letters  $R$  and  $P$ . Taking into account the selection rules discussed above, we obtain the following frequency formula:

$$\nu = \nu_{vib} \pm \frac{h}{4\pi^2 I} j \quad (j = 1, 2, \dots).$$

The plus sign corresponds to transitions to higher rotational levels and the minus sign to lower rotational levels.

This is shown in Fig. 229, which illustrates the spectral band of HCl. The point  $O$  corresponds to  $\nu_{vib}$ , and the vertical lines to the right and to the left indicate

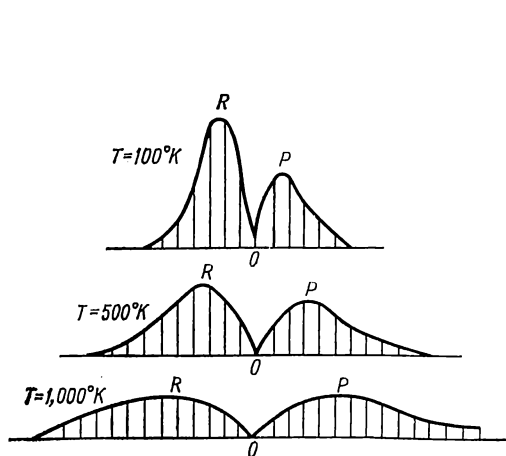


Fig. 229

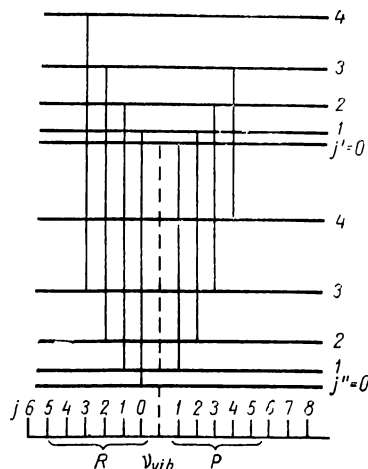


Fig. 230

the obtained frequencies. The height of a line is proportional to the intensity at the given frequency. When the resolution is high, each line appears distinct. On the other hand, when the resolution is low, the lines merge into a band the intensity dependence of which is given by the envelope of the spectral lines. In Fig. 230, we see a diagram of the energy transitions which produce this band. It should be noted that a pure vibrational transition (from  $j' = 0$  to  $j'' = 0$ ) is forbidden and as a result there is a gap in the middle of the band. There is an absorption maximum to the right as well as to the left of the vibration frequency. For the reason discussed in the preceding article, the absorption maxima occur for the  $j$ -values which are most frequently encountered at the given temperature. Therefore, as the temperature increases, the shape of the spectral band changes as shown in the diagram.

**Vibrations of a Polyatomic Molecule.** A polyatomic molecule may execute a large number of vibrational motions. This number is equal to the number of vibrational degrees of freedom of the molecule and may be calculated as follows.



A molecule consisting of  $N$  atoms has  $3N$  degrees of freedom. Three of them are associated with the coordinates of the molecule's centre of mass. In the general case, the number of rotational degrees of freedom is also equal to three. But linear molecules have only two rotational degrees of freedom since rotation about a line passing through the centres of the atoms is physically meaningless. Thus, the number of vibrational degrees of freedom and, hence, the number of vibration frequencies is equal to  $3N - 6$  or  $3N - 5$ . If the dipole moment of the molecule does not change for a given vibration, the corresponding frequency will not be manifested. (We shall return to the problem of so-called inactive vibrations later.) Be that as it may, the number of vibration frequencies and, hence, the number of bands in the infrared spectrum, is strictly determined by the number of atoms in the molecule and by its symmetry.

In the absence of molecular rotation, i.e., in the case of solids, an infrared absorption spectrum consists of lines which correspond to vibrational transitions.

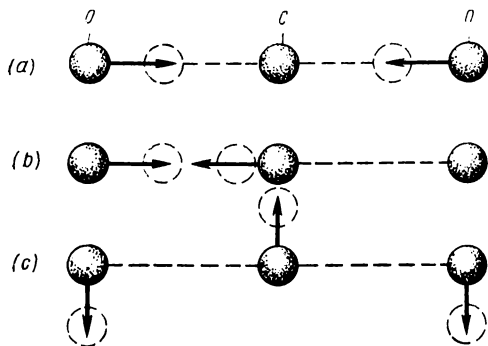


Fig. 231

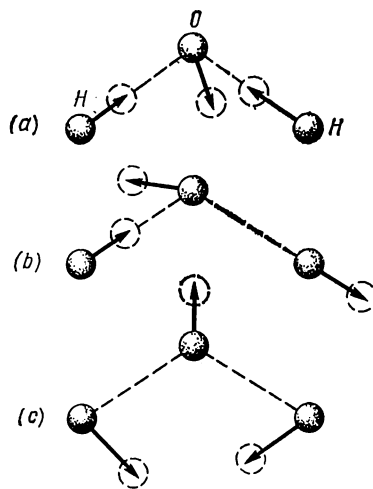


Fig. 232

Since, perforce, thick layers are used in such an investigation, there is considerable absorption under normal conditions and the lines merge into a band. In liquids, molecular rotation is retarded and the rotational structure of such a band will be smeared, i.e., individual lines can no longer be detected.

Now, let us consider the physical meaning of vibrations in a polyatomic molecule. Actually, what kind of vibrations occur? In the case of a diatomic molecule the situation was clear, i.e., we were dealing with vibrations along a bond line. What quantities vibrate harmonically in polyatomic molecules?

For any molecular vibration, the deviations of atoms from their equilibrium positions may be described by displacements along a bond and by the distortion of bond angles. The instantaneous configuration of a vibrating molecule may be completely described by  $(3N - 6)$   $q_i$  coordinates (using the word coordinate in its broad sense). For an arbitrary choice of the  $q_i$  coordinates, their values will not obey a simple vibration law. The law of change with respect to time of each  $q_i$  can be represented by a complex, albeit periodic, curve. However, it turns out that it is possible to describe a vibrating molecule by  $(3N - 6)$   $Q_i$  numbers which

vary harmonically with frequencies  $\nu_i$ . These  $Q_i$  "coordinates" are called *normal coordinates* and the frequencies  $\nu_i$  are called *normal vibration frequencies*.

The fact that it is possible to introduce normal coordinates means that the periodic curves of change of any  $q_i$  coordinates may be resolved into spectra of normal vibration frequencies. We can always assume that a vibration spectrum consists of normal vibration frequencies.

What is the nature of  $Q_i$  coordinates? Are they obtained only for a particular choice of coordinate system? The answer to the latter question is no. First and foremost, normal coordinates are linear combinations of  $q_i$  displacements. Therefore, a normal coordinate describes the vibration of a molecule as a whole. Examples of normal vibrations are illustrated in Figs. 231 and 232 for  $\text{CO}_2$  and  $\text{H}_2\text{O}$  molecules. The actual vibration of a molecule is the resultant of the indicated motions.

The normal vibration frequencies of a molecule can be determined from its spectrum. These can then be used to obtain a clear picture of the molecular vibrations.

The characteristic nature of many vibration frequencies is of great practical importance. Careful study has shown that basically in certain normal vibrations only one interatomic spacing or one bond angle varies. If a molecule preserves that bond, such a frequency varies little in a group of related compounds. This fact is utilised in chemistry.

The vibration frequencies of a molecule are measured not only by means of infrared absorption spectra but by means of Raman spectra as well. As will be seen below, these two methods effectively supplement each other.

#### Sec. 206. RAMAN SCATTERING OF LIGHT

Raman scattering refers to the particular case of the scattering of light of frequency  $\nu$  by a substance when, in addition to the strong scattering of light of constant frequency  $\nu$ , there appears a series of lines of lower and higher frequencies.

Usually, observations are made at right angles to the incident light. A mercury lamp provides the required radiation. The spectrum of this radiation contains several intense lines, the most important of which is a blue line corresponding to a wavelength of  $4,358 \text{ \AA}$ . By means of a spectrograph, one can obtain a photograph of the scattered radiation spectrum. Such a photograph is shown in Fig. 233. The main characteristic of such a spectrum is the following. About each excited line there appear identical groups of considerably weaker lines. These satellites are usually spaced symmetrically to the right and to the left, but they may differ in intensity. This phenomenon was discovered independently by Raman in India and Landsberg and Mandelstam in the Soviet Union. Raman's work, however, was published first\*. Hence, such spectra are called Raman scattering spectra.

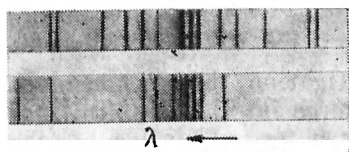


Fig. 233

The spectral pattern can be explained as follows. Basically, a photon  $h\nu$  is scattered by a molecule elastically, i.e., the frequency remains constant. However, in addition to such scattering, it is also possible to have scattering with some loss of energy; such energy may be expended in the transition of a molecule from one

\* Raman sent a telegram about his discovery to the British journal *Nature*.

level to another. Let us assume that a photon  $h\nu$  has lost an amount of energy equal to that required to raise a molecule from the zero vibrational level to the first level. The energy loss is  $\mathcal{E}_{v=1} - \mathcal{E}_{v=0} = h\nu_{vib}$ . Therefore, the scattered photon has an energy  $h(\nu - \nu_{vib})$ . An associated line or "satellite" appears in the spectrum on the side of lower frequencies.

The lower frequencies,  $\nu - \nu_{vib}$ , are called Stokes lines and the higher frequencies anti-Stokes lines. Scattering with a frequency greater than  $\nu$  occurs when a photon hits an excited molecule. In such a case, the photon is scattered, but it simultaneously gains the "extra" energy due to the transition of the molecule to a lower level. If the excited molecule was at the first vibrational level, the photon increases its energy by  $h\nu_{vib}$  and the frequency  $\nu + \nu_{vib}$  appears in the spectrum.

This scattering mechanism excellently explains the difference between the intensities of the red and violet lines. At room temperature, most molecules are at the zero level with an energy  $\frac{1}{2}h\nu_{vib}$ . A smaller number of molecules are at the first excited level with an energy  $\frac{3}{2}h\nu_{vib}$ . Therefore, it is clear that the intensity of the violet lines must be less. Moreover, at low temperatures the violet lines practically disappear. The ratio of the intensities of the violet lines to the red lines is proportional to the ratio of the number of atoms in the first state to the number in the zero state. According to the Boltzmann law,

$$\frac{N_{v=1}}{N_{v=0}} = \frac{e^{-\frac{3}{2} \frac{h\nu_{vib}}{kT}}}{e^{-\frac{1}{2} \frac{h\nu_{vib}}{kT}}} = e^{-\frac{h\nu_{vib}}{kT}}$$

This formula agrees excellently with experimental results.

Thus, the displaced lines of a Raman spectrum are shifted by amounts equal (in energy units) to the difference between the energy levels of the given molecule.

We have discussed the case of two vibrational levels, but it is clear that the discussion is valid for all energy transitions—pure rotational, vibration-rotational, etc. The satellite lines closest to the main scattering line correspond to the lowest energy transitions. The rotational spectrum is located considerably closer to the main line than the vibration-rotational spectrum.

In the case of absorption spectra, the selection rules for the oscillation quantum number are the same as those for Raman scattering, but the selection rules for the rotational quantum number are different. Transitions are allowed for which

$$\Delta v = \pm 1 \quad \text{and} \quad \Delta j = 0, \pm 2.$$

Thus, the vibration-rotational band consists of a pure vibrational line displaced from the excited line by  $\nu_{vib}$  and a series of lines displaced from the excited line by  $\nu_{vib} - 2\nu_{rot}$  and  $\nu_{vib} + 2\nu_{rot}$ .

Raman spectra are usually obtained by scattering from liquids. The lines  $\nu_{vib} \pm 2\nu_{rot}$  appear smeared, but the lines of the pure vibration spectrum are distinct.

Raman spectra have an important advantage over infrared spectra. The measurements are transferred, so to speak, to the visible region. The frequencies which were measured directly in the infrared spectrum are determined as the difference between the main line and the Raman line with approximately the same accuracy.

It would seem that one could dispense with the infrared spectrum. However, this is not always the case. In certain respects, the infrared and Raman spectra supplement each other.

What is the difference between the process of wave radiation from a molecule and the process of scattering from a molecule? In both cases a molecule sends wavelets into space, i.e., in both cases a molecule behaves during radiation like a dipole. However, in the first case a molecule behaves like a dipole in the absence of an external field, while in the second it behaves like a dipole when acted upon by the field of an incident wave. Thus, radiation or absorption will occur when changes in the state of a molecule (vibration, rotation, etc.) are accompanied by changes in the induced dipole moment, i.e., in the polarisability  $\beta$ . Theoretical calculations indicate that such a change should occur when the configuration of the molecule passes through equilibrium.

Lines will occur in an infrared spectrum for vibrations which satisfy the condition

$$\left(\frac{d\mu}{dr}\right)_{r=r_0} \neq 0.$$

Raman lines will occur for vibrations which satisfy the condition

$$\left(\frac{d\beta}{dr}\right)_{r=0} \neq 0.$$

Quite often these conditions exclude each other. Therefore a certain vibration may be active in the infrared spectrum but inactive in the Raman spectrum, and vice versa.

A  $\text{CO}_2$  molecule may serve as an example. One of such a molecule's three vibrations—the linearly symmetrical vibration—leaves the dipole moment unaltered and equal to zero. That vibration is inactive in the infrared spectrum. In the Raman spectrum, on the other hand, only that vibration will be active; the other two will be absent. In the case of an anti-symmetrical vibration, one may reason as follows: in both extreme positions, the deformation of the electron cloud, and hence the polarisability is the same. During vibration the polarisability changes in the same manner in both half-periods, and at the equilibrium position passes through a minimum or maximum, but this does not mean that

$$\left(\frac{d\beta}{dr}\right)_{r=0} = 0.$$

We shall not discuss these regularities any further. They have been studied in detail and the results are available in tabular form. Such tables enable us to determine from the symmetry of a molecule the number of vibration frequencies in its infrared and Raman spectra. The converse is also true, namely, the symmetry of a molecule can be determined from the number of lines in its spectra.

#### Sec. 207. ABSORPTION SPECTRA

Let us consider absorption spectra in which the transitions are in the visible and ultraviolet regions. The magnitude of a photon of the incident light will be of the same order of magnitude as the difference between the electronic levels of a molecule. Electron transitions become possible. However, as has been indicated already, electron transitions are accompanied by changes in vibrational and rotational energy. Therefore, a very broad band is associated with every such transition. Moreover, under normal experimental conditions this band is continuous, i.e., its "vibration-rotational" structure is not discernible. Each electron transition band contains numerous narrow vibration-rotational transition bands, whereby all changes of the oscillation quantum numbers are possible.

The properties of a molecule in an excited electronic state differ from its properties when it is at a zero electronic level. When a molecule becomes excited, the system of vibrational and rotational levels and hence the vibration frequencies, i.e., the differences between the vibrational levels, change. Also, the shape of the potential curve and the equilibrium spacings between atoms change.

Absorption curves in the visible and ultraviolet regions are sufficiently characteristic to be used in identifying substances.

The dependence of the absorption of light on the thickness of a layer of substance may be expressed as follows (cf. p. 89):

$$I = I_0 e^{-\mu x},$$

where  $I_0$  is the intensity of the incident beam,  $I$  the intensity of the transmitted beam,  $x$  the thickness of the layer, and  $\mu$  the absorption coefficient for light.

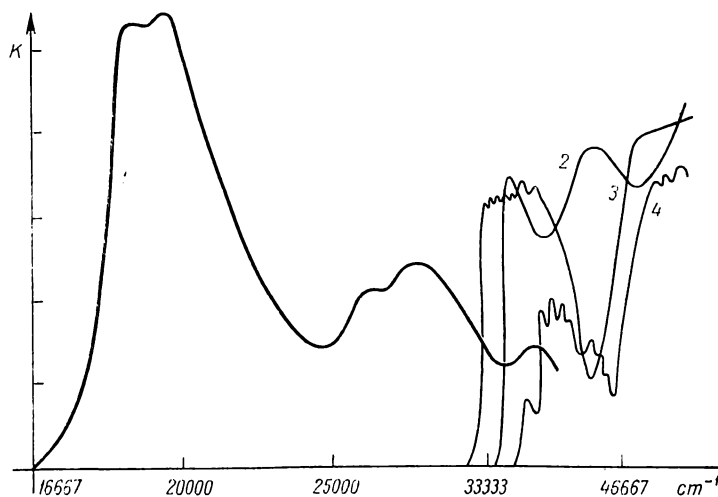


Fig. 234

The value of the absorption coefficient depends on the wavelength of the incident light. A curve of  $\frac{I}{I_0}$  as a function of wavelength is sometimes called an absorption curve. Usually, however, this term refers to a curve of  $\mu$  as a function of  $\lambda$  or  $\nu$ .

The relation for the absorption of light in solutions may be written in the form

$$I = I_0 e^{-kNl} \quad \text{or} \quad I = I_0 e^{-\epsilon cl},$$

where  $l$  is the distance traversed by the beam in the substance and  $kN$  and  $\epsilon c$  are expressions for the absorption coefficient of the solution. It is quite reasonable to assume that the absorption coefficient is proportional to the concentration of the substance, which may be expressed as the number of molecules  $N$  per unit volume or the number of moles of substance  $c$  per litre of solution. In the case of solutions, the term "absorption curve" usually refers to a curve showing the dependence of the coefficient  $k$  or  $\epsilon$  on  $\lambda$ .

Examples of absorption curves in the visible and ultraviolet spectrum are shown in Fig. 234. Curve 1 is for Congo red, 2—for aniline, 3—for phenol and 4—for benzene.

## Sec. 208. MAGNETIC RESONANCE \*

Let a substance containing particles of spin  $s$  and magnetic moment  $M$  be placed in a constant magnetic field of intensity  $H$ . The potential energy of such a particle is equal to the scalar product  $\mathbf{MH} = M_z H$ . According to the general law of quantum mechanics, this energy can attain only a discrete set of values corresponding to  $2s + 1$  possible orientations of spin and magnetic moment in space.

How can we detect the resulting system of energy levels? As usual, it is determined by the energy transitions. The selection rules allow transitions between neighbouring levels only if the difference between their  $s$  values is equal to one. Let, for instance, in one state

$$M_z = g\mu s,$$

and in another

$$M_z = g\mu (s - 1).$$

Hence, the difference in energy is

$$\mathcal{E}_2 - \mathcal{E}_1 = g\mu H.$$

The energy levels will be equally spaced.

To the computed difference in levels there corresponds the frequency of a radiated or absorbed quantum of energy equal to

$$\nu = \frac{g\mu}{h} H.$$

For an electron

$$\nu = 2.8 \times 10^6 H,$$

and for a proton

$$\nu = 3.46 \times 10^3 H.$$

We see that for each value of  $H$  there is a corresponding characteristic frequency known as *the magnetic resonance frequency*. For the realisable range of field intensities, these frequencies lie in the radio band: in the case of nuclei, in the short and ultrashort wavelength region; and in the case of electrons, in the centimetre wavelength region.

Experiments and theory show that in practice it is not possible to detect the radiations corresponding to these frequencies. But it is possible to successfully observe resonance absorption of electromagnetic waves of corresponding wavelength. For this purpose, the substance is placed in a coil connected to a high-frequency generator and the coil is then located in a constant magnetic field. The resonance may be "trapped" by varying the field intensity while the frequency is kept constant, or by changing the frequency while  $H$  remains unchanged. Magnetic resonance is extremely selective. The width of the absorption peak is of the order of 0.1 MHz at 460 MHz.

The method of magnetic resonance is widely used in studying various substances. Detection of electronic as well as nuclear resonance is of great interest. The presence of electrons having uncompensated spins indicates to the chemist that so-called free radicals are present and enables him to determine the character of chemical bonds. Nuclear resonance makes it possible to determine the chemical composition of a substance. But the following important fact should be noted. The magnetic resonance effect is so sensitive as to enable the detection of the superposition of the electron shell field of an atom on the external field. It turns out that the nature of this "supplementary" field depends on the properties of

---

\* Prior to studying this and next sections, read Section 214.

the chemical bond between a given atom and the rest of atoms. Thus, the resonance frequencies of a given atom slightly vary depending on its chemical bond. This phenomenon is known as *chemical shift*.

Figure 235 represents an oscillogram of the absorption spectrum of a chemical compound. It illustrates magnetic resonance for fluorine nuclei. We see four peaks, one of which is three times as high as the other three. In the molecule whose structural formula is shown in the figure, four "different" fluorine atoms are to be seen.

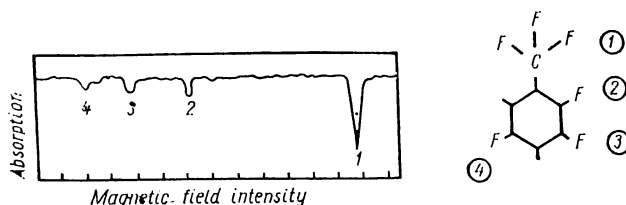


Fig. 235

There are three times as many fluorine atoms in the  $\text{CF}_3$  group as each of the other three "chemically different" fluorine atoms in the molecule. Chemical displacement has separated the nuclear resonances of the fluorine atoms and made it possible to determine the structural formula of this compound.

Thus, the phenomenon of nuclear resonance provides us with a new method of chemical analysis. Instead of obtaining only a gross chemical formula (in our example, the total number of fluorine atoms relative to, say, hydrogen atoms), we can now obtain a detailed picture of a chemical formula, i.e. the proportions of differently bound atoms of a single kind.

#### Sec. 209. QUADRUPOLE RESONANCE

The above discussed scheme of molecular energy levels lacks certain detail. It turns out that each rotational level has an infrastructure. Between the electron shell of a molecule and the atomic nuclei there may exist yet another interaction: an atomic nucleus may possess an electric quadrupole moment and depending on the orientation of the atomic nucleus relative to the electronic shell the molecule may possess different values of energy. The magnitudes of the energy associated with this interaction are quite small and the corresponding energy levels are grouped around the rotational levels. Thus, in order to characterise a molecule, one must specify its quadrupole energy level in addition to its electronic, vibrational, and rotational levels.

Quadrupole interaction does not always exist. If it does, the rotational transitions discussed above are in fact rotational-quadrupole transitions. Pure quadrupole transitions, i.e., transitions between separate quadrupole levels, may be observed. Moreover, rotational-quadrupole transitions may be resolved. Both problems can be solved by radio-spectroscopic methods. Pure quadrupole transitions lie in the range of 1-800 MHz, i.e., in the short-wavelength band. Rotational-quadrupole transitions may be detected by studying the absorption of microwaves (millimetre waves) in gases.

Pure quadrupole transitions are of primary interest. They can be observed in solids and in certain liquids.

The following formula gives the energy of interaction between an atomic nucleus and the electronic cloud of a molecule for the case of an axially symmetrical

field (such a field exists in all linear molecules):

$$\xi = eQq \frac{3m^2 - s(s+1)}{4s(2s-1)}$$

where  $Q$  is the quadrupole moment of the nucleus and  $q = \frac{\partial^2 V}{\partial z^2}$  is the second derivative of the electric potential along the axis of symmetry of the field. This effect does not occur in the case of nuclei having a spin of 0 or  $1/2$ . Also, no interaction occurs when the electron cloud surrounding a nucleus is spherically symmetrical.

There is a limited number of levels. If the selection rules are taken into consideration, one finds that the number of possible transitions becomes quite small. For instance, one line occurs when  $s = 1$  and  $s = \frac{3}{2}$ , two lines when  $s = \frac{5}{2}$ , and three lines when  $s = \frac{7}{2}$ .

A quadrupole absorption spectrum can be observed using a generator whose frequency is continuously varied in the wavelength interval under investigation. The resolving power of radiospectroscopic methods is very high. In the case of a spectral line of the order of 30 MHz, the width of the line is equal to several hundred cycles per second.

The electric quadrupole moment  $Q$  of the nucleus is a constant of an atomic nucleus. It describes the deviation from spherical symmetry of the distribution of electric charge in a nucleus. The value of  $Q$  in square centimetres can be determined experimentally from the above formula if  $q = \frac{\partial^2 V}{\partial z^2}$  is known and the quadrupole frequencies can be measured. The deviation from spherical symmetry of the distribution of electric charge in a nucleus can be determined to a first approximation by representing the nucleus as an ellipsoid of revolution. If the nucleus is elongated in the spin direction, then  $Q > 0$ , and vice versa.

An ellipsoidal nucleus tends to become oriented in a very definite manner in the field of an electron shell. The main energy level corresponds to an arrangement in which the axis of symmetry of the field and the axis of the ellipsoid coincide. Due to the discreteness of energy, in excited states the axis of the ellipsoid can assume only several selected orientations with respect to the axis of symmetry of the field. The energies of these quantum states are just computed by the above formula. An electromagnetic wave impinging on a molecule is absorbed if the magnitude of the photon corresponds to the energy of transition from one orientation of the ellipsoid to another.

The study of quadrupole spectra began quite recently. They are of great scientific interest since such spectra enable us to measure frequencies with extremely high accuracy and the quadrupole frequency responses to very small changes in the electric field of the molecule of which the nucleus is a component as well as of neighbouring molecules.

Suffice it to say that quadrupole frequencies will differ noticeably in crystalline varieties of one and the same substance. Thus, a nucleus reacts not only to changes in the field due to close electrons, but also to changes in the field due to electrons situated far away.

**Example.** The electric quadrupole moment of a nucleus of  $\text{Cl}^{35}$  is  $Q = -0.07 \times 10^{-24} \text{ cm}^2$ . Quadrupole resonance occurs in  $\text{Cl}_2$  at a frequency of  $\nu = 54.47 \text{ MHz}$ . For a nucleus of  $\text{Cl}^{35}$ , the spin  $s$  is equal to  $\frac{3}{2}$ . This means that the quantum number  $m$  attains the following values:  $\frac{3}{2}, \frac{1}{2}, -\frac{1}{2}$  and  $-\frac{3}{2}$ . Since quadrupole interaction energy is a function of  $m^2$ , when the energy



quantum  $h\nu$  is absorbed, only one transition is possible: from the level corresponding to  $|m| = \frac{3}{2}$  to the level corresponding to  $|m| = \frac{1}{2}$ .

The resonance condition is:

$$h\nu = \mathcal{E}_1 - \mathcal{E}_3 = e\eta Q \left[ \frac{3\left(\frac{1}{2}\right)^2 - \frac{3}{2}\left(\frac{3}{2}+1\right)}{4 \times \frac{3}{2}\left(2 \times \frac{3}{2} - 1\right)} - \frac{3\left(\frac{3}{2}\right)^2 - \frac{3}{2}\left(\frac{3}{2}+1\right)}{4 \times \frac{3}{2}\left(2 \times \frac{3}{2} - 1\right)} \right] = -\frac{e\eta Q}{2}.$$

Measuring the resonance frequency  $\nu$  and knowing from other data the value of the quadrupole moment  $Q$  of a  $\text{Cl}^{35}$  nucleus, we find the gradient of the electric field created by electrons at the centre of a  $\text{Cl}^{35}$  nucleus in a  $\text{Cl}_2$  molecule:

$$\eta = \left| \frac{\partial E}{\partial z} \right| = \frac{2h\nu}{eQ} = \frac{2 \times 6.6 \times 10^{-27} \times 54.5 \times 10^6}{4.8 \times 10^{-10} \times 0.07 \times 10^{-24}} = 2.14 \times 10^6 \text{ CGS units.}$$

## Sec. 210. GAS LASERS

Lasers (for Light Amplification by Stimulated Emission of Radiation) or stimulated radiation oscillators represent non-equilibrium (unbalanced) systems with inverse population of energy levels intended for obtaining powerful light fluxes.

The particles of an atomic or a molecular gas which is in a state of thermal equilibrium are distributed among the energy levels in accordance with the Boltzmann law, i.e., the number of particles located on a higher energy level  $E_2$  is  $e^{(E_2-E_1)/kT}$  times less than the number of more stable particles with the energy  $E_1$ . In the normal state, energy levels are populated in such a way that the higher a level is, the less the number of particles it contains.

But this is true for a gas which is not supplied with energy. But if a gas discharge takes place, then the situation is changed. The particle distribution may not only disobey the Boltzmann law, but also the case of *inversion* of the normal population distribution is possible. This means that the upper energy levels turn out to be more populated than the lower levels. If such a situation is achieved by energy supply (by what is called pumping), a laser design becomes possible.

It is clear why population inversion is a necessary condition for creating a laser. First of all, we are not interested in spontaneous radiation. As it was mentioned above, spontaneous radiation is nondirectional and incoherent. Hence, the topic of discussion is stimulated (or induced) emission of radiation. Since the probabilities of a transition of a particle hit by a photon upward and downward are equal, amplification by stimulated emission of radiation becomes possible only if the upper energy level has a greater concentration of atoms than the lower one has.

The situation is illustrated by the diagram given in Fig. 236. In an unexcited state the particles of the lasing medium are located mainly on the lower energy level (Fig. 237). When pumping begins, the desired population inversion takes place (Fig. 238). A certain particle spontaneously emits a photon which is capable of stimulating the emission of other particles. This activity of the photon is continued until it is absorbed by a particle found on the lower energy level (Fig. 239).

It is possible to realize lasers operating under pulse conditions: by pumping, the particles are brought to the upper energy level, and then, during a very short interval of time, this reserve of energy is completely given up in the form of stimulated emission of radiation.

Gas lasers operate in a continuous duty. To make it possible, we must have available a system of particles which possess the following peculiarities. The pumping must transfer the particles from the ground state to the upper laser

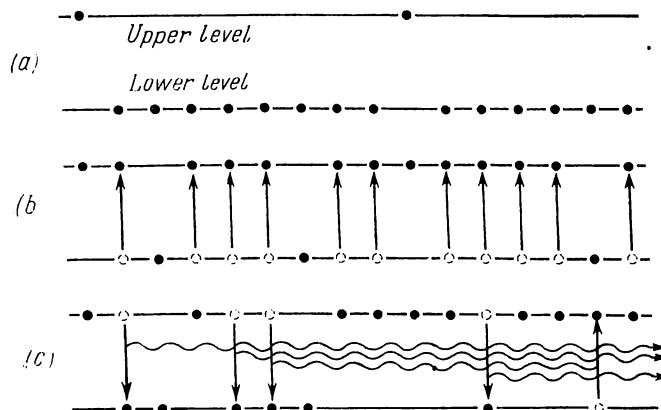


Fig. 236

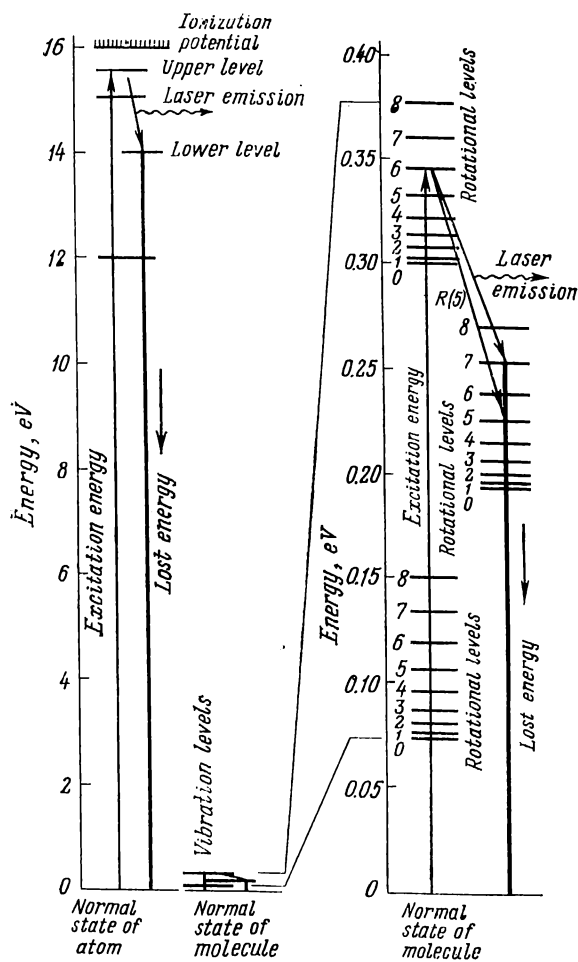


Fig. 237

level. Laser emission consists in the transition of particles from the upper laser level to the lower one. From the lower laser level the system transfers to the ground state through spontaneous emission.

It is clear from what was said that excitation must not transfer the particles to the lower level. Besides, the lower level must quickly get freed, i.e., the lifetime in this state must be essentially less than the lifetime on the upper laser level.

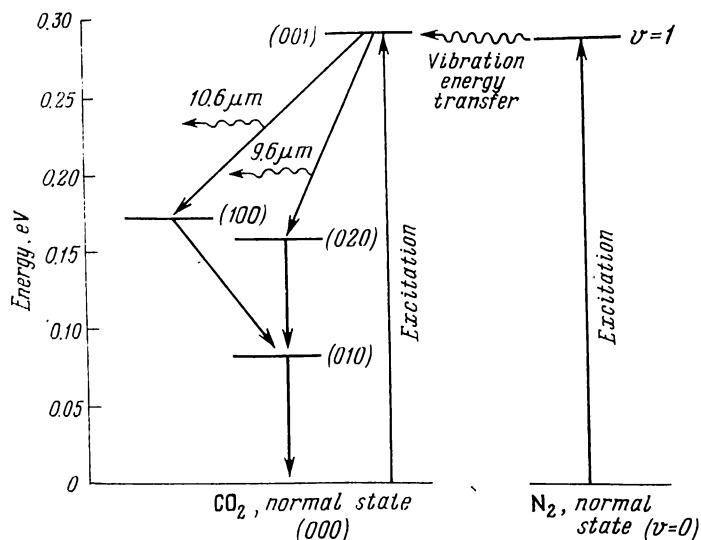


Fig. 238

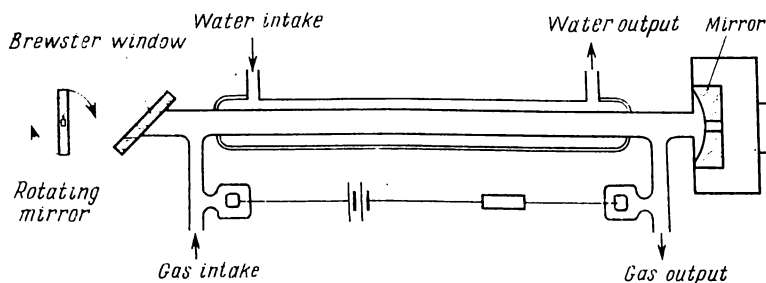


Fig. 239

The transition of a particle from the lower level to the ground state does not contribute to the laser emission. This is an inevitable loss and it can be lessened only by one method, that is by selecting systems in which the difference between the energies of the upper and lower levels is great as compared with the difference in the energies of the lower and ground levels.

The ratio of the energy of an emitted photon to the energy of excitation is known as the *absolute maximum efficiency* of a laser. Of course, it is much less than actual efficiency, since the pumping energy is inevitably spent not only on raising a particle to the desired upper level.

In one and the same gas there may exist several possible upper and lower laser levels. The conditions favourable for producing photons of a certain energy, i.e. for separating out two levels as the upper and the lower laser levels are obtained

by designing the laser like a resonating hollow. If it represents a gas-discharge tube with mirrors fitted at the base of the column, then by means of a micrometer screw the desired photons are selected by varying the length of the column. Since waves are reflected from the mirrors several times, favourable conditions are established only for the light the integral number of wavelengths of which is exactly equal to the length of the column. Gas lasers in which energy is pumped by means of an electric discharge were designed almost for all the elements. Laser emission was obtained with wavelength from 0.2 to 133  $\mu\text{m}$ .

Widely used are lasers whose medium is a mixture of neon and helium. In most cases, a near-infrared light is created with wavelength of 1.13  $\mu\text{m}$ .

Gas mixtures are used in lasers for the following reason: in some cases, using a gas discharge, it is simpler to excite particle *A* which will transfer the excitation to particle *B* than directly excite particle *B*.

We shall describe here in more or less detail the operation principle of the most powerful modern lasers, and namely, the laser operating on carbon dioxide gas.

The main idea of utilising molecular gases consists in the possibility to essentially increase maximum efficiency, using the vibrational levels of the ground electronic state as the upper and lower laser levels. A comparison in this respect to an atomic and molecular lasers is given by the diagram of Fig. 237.

Various vibrations of the molecule  $\text{CO}_2$  were discussed in one of the previous sections (see Fig. 234). Any vibrational state is characterised by three quantum numbers  $\nu_1\nu_2\nu_3$ ,  $\nu_1$  referring to a symmetrical vibration (*a*),  $\nu_2$  to vibration (*c*), and  $\nu_3$  to a linear asymmetrical vibration (*b*).

First of all, the researcher must find out the lifetimes of a molecule in different states. These lifetimes may differ by several orders. Further, the probabilities of transitions on this or that level under the action of electron knocks are extremely essential.

It turns out that, from all viewpoints, level 001 is suitable as the upper laser level and 100 or 020 as the lower level. From these levels a molecule transfers to level 010, and then returns to the ground state. The diagram of these transitions is shown in Fig. 238; for the sake of obviousness, the vibrational levels are not shown. As is seen from the diagram, the laser has a high maximum efficiency of 40 and 45 percent for radiations of 10.6  $\mu\text{m}$  and 9.6  $\mu\text{m}$ .

Both absolute and practical efficiencies of this system are high, since the gas-discharge electrons transfer the molecules mainly on levels  $00\nu_3$ . A remarkably convenient circumstance is that excitations on any level  $00\nu_3$  are equally suitable. Recall that the vibrational levels are equidistant. Therefore a collision of molecules in states  $00\nu_3$  and 000 yields molecules in states  $00(\nu_3 - 1)$  and 001. That is, in the final run, there occur desired molecules situated on the upper laser level.

Despite the fact that excitations to any of the levels yield a positive contribution to the laser operation, nevertheless, the electrons spend a greater energy on ion excitations of molecules. Excitation becomes considerably more selective when nitrogen molecules are added to  $\text{CO}_2$ .

Nitrogen has a vibrational level for  $\nu = 1$  of the ground electronic state with an energy value equal to energy 001 of the molecule  $\text{CO}_2$ . This excited state has quite a long life, and the nitrogen molecule descends to the zero level mainly in one way, by giving up its energy to the molecule  $\text{CO}_2$  in state 000 (see the diagram of Fig. 238). The equidistant vibrational levels of nitrogen make all of its vibrational states (of the ground electronic level) effective.

Considerations like those described above are not quite rigorous since too many factors affect the practical efficiency of a laser. However, they demonstrate the methodology of searching for lasing media. In the final run, everything is solved

by experience. If, say, the lower laser level is freed too slow, then it is advisable to add other gases. The search for such admixtures is mainly of an empiric character, but the results may turn out to be very significant. For instance, at a gas pressure of 1 mm Hg the molecules  $\text{CO}_2$  without contaminants undergo about 100 level-liberating collisions per second. In the presence of helium and water the corresponding figures are 4,000 and 100,000, respectively.

Until now, nothing was said about the effect of the rotational structure of vibrational levels on the power of a  $\text{CO}_2$  laser. If the transitions among all the rotational sublevels took part in emission of the laser, then the emission would not be strictly monochromatic. By making use of one delicate effect we are going to describe we may succeed in making the laser operate on transition between certain sublevels. The 20th level of the *P*-branch of the transition (004)-(100) is usually used, which gives a pencil of photons with a wavelength of 10.5915  $\mu\text{m}$ .

At a room temperature the average kinetic energy of the molecule  $\text{CO}_2$  is equal to 0.025 eV. The distance between vibrational levels is greater than this magnitude, and the distance between rotational levels is less than this value. That is why the transitions of molecules due to thermal collisions from one rotational level to another occur much more frequently (10 million per second) than the transitions between vibrational levels (1000 transitions per second). Accordingly, the lifetime of the vibrational state is equal to  $10^{-3}$  s, and that of the rotational state to  $10^{-7}$  s. Thus, inside rotational "stories" of each vibrational level there succeeds to set in Boltzmann distribution corresponding to thermal equilibrium.

Under these conditions, it is sufficient to set the column for a certain transition to make it predominant. Indeed, suppose we choose the transition *P* (22), i.e. the transition from the 21st rotational sublevel (001) to the 22nd sublevel (100). As the 21st sublevel becomes deserted, it is gradually occupied by other molecules coming from other rotational sublevels, and, thus, the Boltzmann distribution is maintained. This is how a transition put in favourable conditions wins the competition with other possible transitions.

This characteristic property of the  $\text{O}_2$ -laser determines its great merits, making it highly monochromatic and enabling the operator to vary, though in a limited range (9-11  $\mu\text{m}$ ), the wavelength of stimulated radiation.

A long lifetime of vibrational states makes it possible to switch the  $\text{CO}_2$ -laser to pulse conditions. For this purpose, one of the two stationary mirrors are replaced by a rotating mirror. The laser is set in operation each time the rotating mirror occupies the proper position relative to the fixed mirror.

Rated at a constant power of 50 W, under pulse conditions the device is capable of yielding 50 kW by flashes of duration 150 ns at a rate of 400 flashes per second.

The  $\text{CO}_2$ -laser is usually manufactured in the shape of a tube 2 m long through which a gas flux is passed. Its diagram is shown in Fig. 239.

The possibilities of the described laser are almost fantastic. Its coherent infrared radiation focused at an area of 0.001  $\text{cm}^2$  yields an intensity of  $10^6$  W/ $\text{cm}^2$  under constant conditions and  $10^9$  W/ $\text{cm}^2$  under pulse conditions. The keen laser ray capable of propagating over great distances burns through wood in no time and penetrates through steel during some seconds.

The laser ray can create fields of the order of  $10^6$  V/cm which radically change the properties of a substance.

The creation of lasers gave rise to a number of entirely new physical and engineering researchers. Of great interest is a study of interaction of light with a wavelength of 10-11  $\mu\text{m}$  with semiconductors which are transparent for this range of spectrum. Some applications of lasers were dealt with in the previous sections of the present book.

# The Atomic Nucleus

## Sec. 211. EXPERIMENTAL METHODS OF NUCLEAR PHYSICS

Investigation of the atomic nucleus is inseparably connected with the study of spontaneous or induced decay of atomic nuclei and nuclear particles. By studying the fragments of a disintegrated atomic nucleus and tracing the fate of these particles, we get the possibility of drawing certain conclusions about the structure of the nucleus and about the nuclear forces.

The spontaneous decay of nuclei, that is natural radioactivity, was the first to be studied in detail. Concurrently, physicists began to study cosmic rays—radiation coming to us from outer space and possessing extraordinary penetrating power. In interacting with matter, cosmic particles behave like projectiles. For a long time, cosmic ray investigations were the primary means of studying transformations of elementary particles and to a certain extent of studying the atomic nucleus. At present, investigators concentrate on studying the disintegration of the atomic nucleus due to the bombardment by fluxes of particles created in accelerators.

The experimental methods to be discussed below are applicable to the study of cosmic rays and particles occurring as a result of the nuclear bombardment of one or another target.

**Track Chambers.** The Wilson cloud chamber was the first device enabling the researcher to see a particle track.

When a fast particle passes through a chamber containing supersaturated vapour and creates ions along its path, a track is left that is very similar to the “tail” occasionally seen in the sky after a plane has passed. This track is produced by condensation of the vapour. The ions along the path of the particle are centres of vapour condensation; as a result the track is easily detected. The track of the particle can be observed directly or photographed.

The state of the vapour in the chamber can be controlled by varying the volume of the chamber. This is achieved by means of a piston. Rapid adiabatic expansion brings the vapour to a state of supersaturation.

If a track chamber is placed in a magnetic field, the velocity of a particle in the chamber may be determined from the curvature of its path if the ratio  $e/m$  is known or, conversely,  $e/m$  may be determined if the velocity is known (see the formula on p. 339).

The Wilson cloud chamber is not used any longer in modern researches. Since it is filled with gas, collisions occur rarely. The chamber “cleaning” time is too long: photographs may be taken only after an interval of 20 seconds. Lastly, tracks live a life of the order of one second—the fact which may lead to a displacement of pictures.

In 1950 a new design was suggested—a bubble chamber which, today, plays an important role in elementary-particle physics. This chamber is filled with over-heated liquid. A charged particle creates ions which become surrounded with bubbles, and the latter make the particle track visible. In such a chamber we can produce up to ten photographs per second. The main shortcoming of the chamber is the lack of possibility to control its switching on. Therefore, thousands of photographs are often needed to select one which illustrates the phenomenon under investigation.

Of high importance are so-called spark chambers based on a different operating principle. If a high voltage is applied to a plane capacitor, then an electric spark will be seen jumping between the plates. If the gap contains ions, then sparking will occur at a lower voltage. Thus, an ionising particle flying between the plates creates a spark.

In a spark chamber, a particle itself "switches on" high voltage for a millionth of a second. But the advantages concerning the possibility of switching on at a proper instant are weakened by the following drawbacks: the observer sees only the particles forming an angle of not more than  $45^\circ$  with the plates, the track is very short and not all secondary phenomena have time enough to display themselves.

Not long ago, Soviet investigators suggested a new type of the track chamber (a so-called streamer chamber), which has already found wide application. The block diagram of such a chamber is shown in Fig. 240. A particle entering the

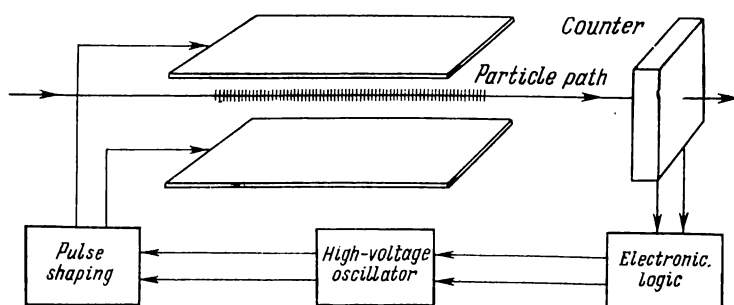


Fig. 240

space between the plates arranged at a greater distance apart than in a spark chamber is detected by a counter. An electronic logical device incorporated in this chamber distinguishes the primary events and chooses one which is of interest for the investigator. At this instant, high voltage is fed to the plates for a short period of time. The ions created along the path of particle motion form streamers which are then photographed. The path of a moving particle is indicated by these streamers. If a photograph is taken along the direction indicated by the streamers, then the particle path looks like a dashed line.

The success of operation of a streamer chamber depends on a correct correlation of the formation of an electron avalanche from a primary ion with the parameters of a high-voltage pulse. In a mixture of 90 per cent of neon and 10 per cent of helium, with the plates 30 cm apart, good results are obtained at a voltage of 600,000 V and pulse duration of  $10^{-10}$  s. In this case, a pulse must be applied not later than in  $10^{-6}$  s after the primary act of ionisation.

A track chamber of this type represents a complex expensive installation which resembles the Wilson cloud chamber not more than a modern particle accelerator does an electron tube.

**Ionisation Counters and Ionisation Chambers.** An ionisation device used in radiation investigations usually represents a cylindrical capacitor filled with gas. The cylindrical plate constitutes one electrode and a thin filament fitted along the axis of the cylinder the other (Fig. 241). The voltage to be applied to the capacitor and the pressure of the gas in the counter must be chosen in a special way, depending on the nature of the problem to be solved. In a widely used modi-

fication of such a device, called *the Geiger counter*, a voltage equal to the breakdown voltage is applied both to the cylinder and to the filament. When an ionising particle enters such a counter through the wall or insulator, there passes through the capacitor a pulse of current which continues to flow until the primary electrons and the created electrons and ions of the self-maintained discharge reach the positive plate of the capacitor. This current pulse can be amplified by ordinary radio-engineering means and the passage of the particle through the counter may be determined by a click or light flash, or by means of a digital counter.

Such a device can be used to count the amount of particles entering the instrument. This requires that the current pulse due to one particle cease by the time

the next particle enters the counter. If the operating conditions of the counter are chosen not in a proper way, the counter begins to "choke" and count incorrectly. The resolving power of an ionisation counter is limited, but still sufficiently high, namely, up to 10,000 particles per second.

We may reduce the voltage so as to obtain such operating conditions under which the current pulse passing through the capacitor may be made proportional to the number of created ions (a proportional counter). For this purpose, it is necessary to operate in a region where the gas discharge is not self-maintained. Primary electrons moving in the electric field of the capacitor accumulate energy, whereupon ionisation by collision commences and new ions and electrons are produced. The first  $n$  pairs of ions produced by a particle entering the counter are transformed into  $kn$  pairs of ions. When the operating conditions are such that the discharge is not self-maintained, the amplification factor  $k$  is constant. Thus, a propor-

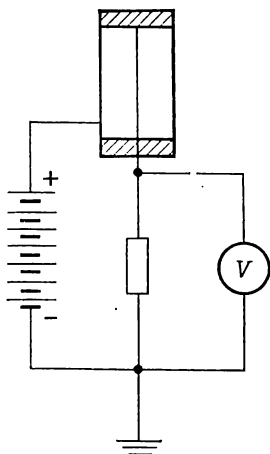


Fig. 241

tional counter not only indicates that a particle has passed through the counter, but also measures the ionising power of the particle.

Just like in the above described Geiger counter, discharge in a proportional counter ceases when no more ionisation takes place. The distinguishing feature of a Geiger counter consists in that a particle entering it behaves like a trigger mechanism, and the breakdown time is independent of the primary ionisation.

Since a proportional counter is sensitive to the ionising power of a particle, the operating conditions of the counter can be chosen so that only certain kinds of particles are recorded by the instrument.

If the operating conditions of the instrument correspond to the saturation current (which can be achieved by reducing the voltage), the current flowing through the counter becomes a measure of the radiation energy absorbed in the instrument per unit time. In such a case the device is called *an ionisation chamber*. The amplification factor  $k$  is then equal to unity. The advantage of an ionisation chamber consists in its high stability of operation. The design of an ionisation chamber varies considerably from case to case. Chamber filling, wall material, and number and shape of electrodes usually vary depending on the purpose of the research. Ionisation chambers vary in size from about a cubic millimetre to several hundred cubic metres. Under the action of a constant source of ionisation, currents ranging from  $10^{-17}$  to  $10^{-7}$  A are produced in ionisation chambers.

**Scintillation Counters.** The method of counting elementary particles by the flashes of fluorescent substance (scintillations) was first used by Ernest Rutherford



in his classical investigations of the structure of the atomic nucleus. Modern instruments bear little resemblance to the simple device used by Rutherford.

A particle impinging on a phosphor\* may produce a flash of light. There exist a large number of organic and inorganic substances which are capable of transforming the energy of charged particles and photons into luminous energy. The duration of afterglow in many phosphors is very short—of the order of a thousand millionth of a second. This makes it possible to design scintillation counters with high counting rates. The light yield of a number of phosphors is proportional to the energy of the particles. This enables us to construct counters intended for the estimation of the energy of particles.

In modern counters phosphors are combined with photomultipliers having ordinary photocathodes that are sensitive to visible light. The electric current obtained in a photomultiplier is amplified and fed to a counting device.

The most widely used organic phosphors include anthracene, stilbene, and terphenyl. These chemical compounds belong to the class of so-called aromatic compounds which contain rings of six carbon atoms. For use as scintillators, these substances must be obtained in the form of monocrystals. Since large crystals are rather difficult to grow and since crystals of organic compounds are very fragile, the use of plastic scintillators, i.e. solid solutions of organic phosphors in transparent plastics such as polystyrene and other such high-polymer materials, is of considerable interest. Of the inorganic phosphors use is made of halides of alkali metals, zinc sulphide, and tungstates of alkaline earth metals.

**Cherenkov Counters.** As far back as 1934, Cherenkov showed that when a charged fast particle moves in a perfectly pure liquid or solid dielectric a peculiar luminescence occurs. This luminescence basically differs from fluorescence, which is related to energy transitions in atoms of the substance, and from bremsstrahlung of the continuous X-ray spectrum type. Cherenkov radiation occurs when a charged particle moves with a velocity exceeding the phase velocity of the propagation of light in the dielectric. The basic feature of this radiation consists in that it propagates along a conical surface in the propagation direction of the particle. The angle of the cone is determined by the formula

$$\cos \theta = \frac{v}{V}$$

where  $\theta$  is the angle between the generatrix of the cone and the direction of motion of the particle,  $V$  is the velocity of the particle, and  $v$  is the velocity of light in the medium. Thus, for a medium with a given refractive index  $n$ , there exists a critical velocity,  $V = v = \frac{c}{n}$ , below which no radiation occurs. At this critical velocity, the radiation is parallel to the direction of motion of the particle. For a particle moving with a velocity very close to the velocity of light ( $v = c$ ), the angle of radiation  $\theta = \arccos 1/n$  has a maximum value. In the case of cyclohexane,  $n = 1.437$  and  $\theta = 46^\circ$ .

Theoretical calculations and experiments show that the Cherenkov radiation spectrum is found mainly in the visible region.

Cherenkov radiation is a phenomenon similar to the formation of a bow wave by a moving ship. In such a case the velocity of the ship is greater than that of the waves on the water surface. Figure 242 illustrates how the Cherenkov radiation is formed. A charged particle moves along the axial line and the electromagnetic wave following the particle temporarily polarises the medium at points of the

\* Phosphors is the name given to a large class of substances which, generally speaking, have nothing in common with the chemical element phosphorus.

particle path. All these points become sources of spherical waves. There exists only one angle for which these spherical waves coincide in phase and form a single front.

Consider two points on the path of the charged particle (Fig. 243). Two spherical waves have been created by them: one at the instant  $t$  and the other at the instant  $t + \tau$ . Obviously,  $\tau$  is the time required for the particle to cover the distance between these two points. In order for these two waves to be propagated at an angle  $\theta$  in the same phase, the time of travel of the first beam must be greater than the time of travel of the second by  $\tau$ . The path covered by the particle during the time  $\tau$  is equal to  $V\tau$ , while the distance traversed by the wave during the same time is equal to  $v\tau$ . Hence we obtain the above given formula:  $\cos \theta = \frac{v}{V}$ .

Today, Cherenkov radiation is widely used as a means of registering elementary particles. Counters based on this phenomenon are known as *Cherenkov counters*.

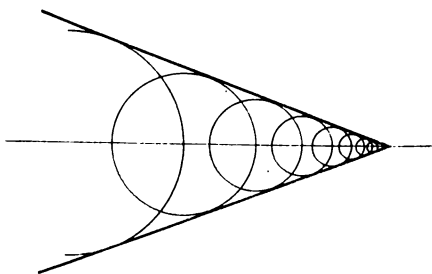


Fig. 242

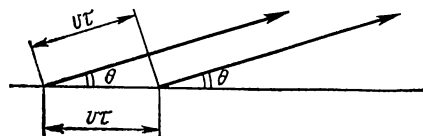


Fig. 243

Like scintillation counters, they contain a luminescent material, photomultipliers, and amplifiers of photoelectric current. Various types of Cherenkov counters have been designed.

These counters have numerous merits and among them are a high counting rate and the ability to determine the charges of particles moving with velocities very close to the velocity of light (it should be mentioned that the light yield sharply depends on the particle charge). Only with the aid of Cherenkov counters it is possible to solve such important problems as the direct determination of the velocity of a charged particle, the determination of the direction in which a high-speed particle moves, etc.

**Arrangement of Counters.** To study transformation and interaction processes of elementary particles, one must be able to detect the emergence of a particle at a particular place and trace its subsequent path. Various arrangements of standard counting circuits are utilised in solving such problems. For example, two or more counters may be electrically connected in such a way that a count occurs only when a discharge begins in all the counters at exactly the same time. This may serve to indicate that a certain particle has passed through all the counters. This type of connection of counters is known as a "coincidence" pattern.

**Nuclear-Emulsion Method.** As is known, a gelatinous film containing microcrystals of silver bromide may serve as the photosensitive layer of a photoplate. Basically, the photographic process consists in the ionisation of these crystals, resulting in the reduction of the silver bromide. This process occurs not only under the action of light, but also may be caused by charged particles. A concealed track is formed in the emulsion when a charged particle passes through it. This track may be seen after the photoplate is developed. Photoemulsion tracks tell

us a great deal about the particles producing them. Strongly ionising particles turn out to leave heavy tracks. Moreover, since the ionisation produced depends on the charge and the velocity of the particle, considerable information may be obtained by simply examining the appearance of the track. Free path of a particle in the photoemulsion is another source of valuable information. By measuring the length of the track, one can determine the energy of the particle.

Ordinary photoplates, having thin layers of emulsion, are hardly suitable for nuclear investigations. Such plates would register only the particles which move strictly along the plate.

Mysovskii and Zhdanov in the Soviet Union, and several years later Powell in Great Britain, introduced the use of photoplates having an emulsion layer 1 mm thick (one hundred times greater than the emulsion thickness in ordinary plates). The complex transformations occurring in the course of particle disintegration process are registered in a clear visual form in the photographic method.

Figure 244 shows a typical photograph obtained by this method. Nuclear transformations have occurred at points *P* and *S*.

In a recent modification of this method, an emulsion chamber of considerable volume is used as the medium for registering the particle track.

**Methods of Analysing Observations.** Using the above described instruments and techniques, a researcher is able to determine the most important constants of an elementary particle, namely, velocity, energy, electric charge, and mass. All these parameters can be determined with a sufficiently high accuracy. Moreover, if a particle flux is available, it is possible to determine the spin and magnetic moment of an elementary particle. This is done by means of a magnetic field used to divide the flux.

It should be remembered that only charged particles can be observed directly. Our knowledge of neutral particles and photons is obtained indirectly, i.e., by studying how these invisible particles affect the charged particles. Nevertheless, our knowledge of invisible particles is highly reliable.

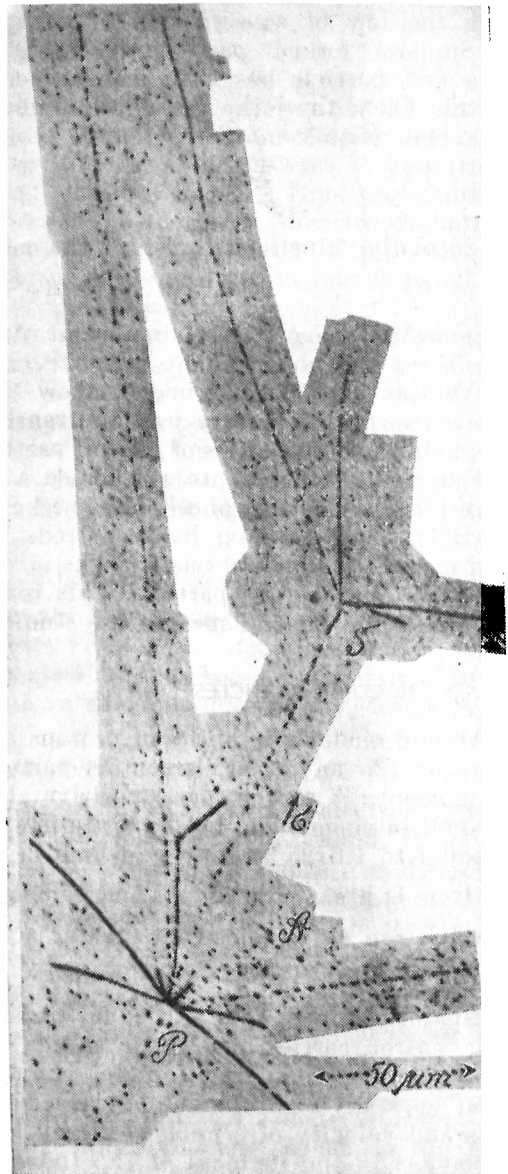


Fig. 244

The laws of conservation of momentum and energy are widely applied in the investigation of elementary particle transformations. Since we are dealing with fast particles, possible changes in mass must be taken into consideration in applying the law of conservation of energy.

Suppose "forked" particle tracks are seen in a photograph. This signifies that the first particle has been transformed into two particles—the second and the third. Then, the following relations must be fulfilled. Firstly, the momentum of the first particle must be equal to the vector sum of the momenta of the created particles:

$$\mathbf{p}_1 = \mathbf{p}_2 + \mathbf{p}_3.$$

Secondly, kinetic energies of the particles must be related as follows

$$K_1 = K_2 + K_3 + \Delta\mathcal{E},$$

where  $\Delta\mathcal{E} = c^2 \Delta m$  (the increment  $\Delta m$  being the difference in mass between  $m_2 + m_3$  and  $m_1$ ).

Nuclear physics experiments show that the laws of conservation are strictly obeyed in all elementary particle transformations. Hence, we may use these laws to explain the properties of neutral particles, which do not leave tracks in a photographic emulsion and do not ionise a gas. When an investigator observes two diverging tracks on a photoplate, he knows that at the branching point a neutral particle transformation has occurred. By determining the momentum, energy, and mass of the formed particles, he may reliably ascertain the values of the parameters of the neutral particle. This is how the neutron was discovered and how neutrinos and neutral mesons are studied (see below).

#### Sec. 212. NUCLEAR PARTICLES

Atomic nuclei are 'built' of protons and neutrons. The basic characteristics of a proton, like any other elementary particle, are its charge, mass, spin, and magnetic moment. A proton has a positive elementary electric charge, i.e., its charge is equal in magnitude, but opposite in sign, to the charge of an electron. Its mass is equal to  $1.6725 \times 10^{-24}$  g, which is 1,836 times greater than the mass of an electron. It has a spin of  $\frac{1}{2}$  and a magnetic moment of  $1.41 \times 10^{-23}$  CGS unit.

A neutron has a somewhat greater mass than a proton, i.e., its mass is equal to  $1.6748 \times 10^{-24}$  g. Its spin is also  $\frac{1}{2}$ . The magnetic moment of a neutron is anti-parallel to the spin and equal to  $0.966 \times 10^{-23}$  CGS unit.

A neutron carries no electric charge and does not leave a track in a Wilson cloud chamber or on a photoplate. The properties of a neutron are mainly investigated by studying collisions between neutrons and various nuclei. Knowing the mass and velocity of a nucleus hit by a neutron, one can determine the neutron velocity  $v_{neut}$  and its mass  $M_{neut}$ . Indeed, according to the laws of elastic impact (see Sec. 16), we get

$$v_{nucl} = \frac{2M_{neut}}{M_{neut} + M_{nucl}} v_{neut},$$

where  $M_{neut}$  and  $v_{neut}$  are unknown quantities. By studying collisions between neutrons and various nuclei, one can roughly determine  $M_{neut}$  if it is assumed that the initial velocity  $v_{neut}$  is the same in different collisions. A precise value for  $M_{neut}$  may be determined from the values of the mass defect of nuclear reactions (see below).

The spin and magnetic moment of neutrons have been directly determined from very interesting measurements on a stream of neutrons passing through magnetised iron. However, we are not going to describe these measurements.

#### Sec. 213. MASS AND ENERGY OF AN ATOMIC NUCLEUS

The number of protons  $Z$  in the nucleus of a given element determines the chemical properties of the element, and the position of an atom in the Mendeleev periodic table is determined precisely by the number of its protons  $Z$ .

A chemical element may have several isotopes, which differ from one another by the number of neutrons in the nucleus. An isotope of a given element is characterised by its mass number  $M$  which is equal to the total number of protons and neutrons in its nucleus\*. Thus, the number of neutrons in a nucleus is equal to  $M - Z$ .

Chemically simple natural substances are a mixture of isotopes. The isotopic composition of a natural substance usually remains unchanged and, therefore, characteristic of a given chemical element. Frequently, one of the isotopes of the mixture predominates. For instance, hydrogen is encountered in nature in the form of ordinary hydrogen  $H^1$  and deuterium  $H^2 = D$ , the percentage of the former being 99.98% and that of the latter 0.02%. The percentage of the isotope  $O^{16}$  in natural oxygen is 99.76%. In natural uranium the percentage of the main isotope  $U^{238}$  amounts to 99.28%.

Let us denote the mass of the isotope  $C^{12}$  of carbon by  $M_0$ . The magnitude  $\frac{1}{12} M_0$  is called an *atomic mass unit*. It is customary to express the atomic weight  $A$  of isotopes and elements in such relative units.

It is determined by precise measurements that the atomic mass unit is equal to a mass of  $1.6604 \times 10^{-24}$  g. The absolute value of the mass of any isotope  $A$  (in grams) is found by the formula

$$M_A = 1.6604 \times 10^{-24} A.$$

The mass of a proton is 1,836 times greater than the mass of an electron. Therefore, the mass of an atom and the mass of its nucleus are almost equal. However, in a number of cases, particularly for light atoms, this difference may be determined by means of precise measuring techniques and should be taken into account. It is obvious that between the mass of an atom  $M_A$  and the mass of its nucleus  $M_N$  there exists the following relation

$$M_N = M_A - Zm.$$

In atomic mass units,  $m = 5.486 \times 10^{-4}$ \*\*. Thus, the difference between  $M_N$  and  $M_A$  is of the order of several hundredths of per cent of these masses (for heavy atoms—of the order of several thousandths of a per cent).

The relative atomic weight of an isotope is close to its mass number, but is not equal to it. In other words, it is approximately equal to the mass number. For example, the mass of  $H^1$  is equal to 1.00807, of  $D^2$  to 2.01463, of  $Ne^{20}$  to 19.9972, etc.

\* The nucleus of an isotope of a given chemical element is denoted by the symbol of the element. The mass number is indicated by a superscript to the right. A subscript to the left is frequently used to indicate the atomic number  $Z$  of the element, but this is not necessary since the chemical symbol determines  $Z$ . For instance, the three isotopes of oxygen may be denoted in two possible ways:  $O^{16}$ ,  $O^{17}$ ,  $O^{18}$  or  ${}_8O^{16}$ ,  ${}_8O^{17}$ ,  ${}_8O^{18}$ . The nuclei of these isotopes contain 8, 9 and 10 neutrons ( $M - Z$ ), respectively.

\*\* This value is obtained by dividing the mass of an electron in grams by  $1.6604 \times 10^{-24}$  g.

The following important conclusion may be drawn from a careful study of a table of isotope masses: the mass of a nucleus is less than the sum of the masses of its constituent particles. For instance, the mass of a neutron is 1.00888 and the mass of a proton is 1.00807; hence the sum of the masses of two neutrons and two protons is equal to 4.0339. But we know that the mass of a helium atom, consisting of two neutrons and two protons, is not equal to this number, being equal to 4.0038. Thus, the mass of a helium nucleus is 0.0301 of an atomic mass unit less than the sum of the masses of the constituent particles of the nucleus. This value is a thousand times larger than the accuracy of measurements.

This difference between the sum of the masses of the constituent particles of the nucleus and the mass of the nucleus is an important example of *mass defect*. Every nucleus has a specific mass defect.

One of the most important conclusions of the theory of relativity is the principle of the equivalence of mass and energy (see Sec. 160). This principle states that if a system acquires or loses a quantity of energy  $\Delta\mathcal{E}$ , the mass of the system

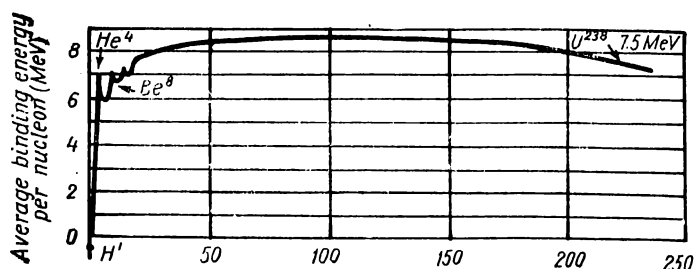


Fig. 245

increases or decreases, respectively, by  $\Delta m = \Delta\mathcal{E}/c^2$ . The mass defect of a nucleus is easily explained with the aid of this principle, i.e. it is a measure of the binding energy of the nuclear particles.

Let us explain what is implied by this statement. In chemistry and physics, binding energy is understood to mean the energy required to completely break a bond. If a nucleus could be divided into its constituent elementary particles, the mass of the system would increase by the value of the mass defect  $\Delta m$ . From the viewpoint of Einstein's law, this means that the nucleus has been given an energy  $\Delta\mathcal{E} = c^2 \Delta m$ , which is just equal to the binding energy. Hence, we find that a change in mass of one atomic mass unit is equivalent to a change in energy of  $1.6604 \times 10^{-24} \times 9 \times 10^{20} \text{ erg} = 1.496 \times 10^{-3} \text{ erg} = 931.8 \text{ MeV}$  ( $1 \text{ eV} = 1.602 \times 10^{-12} \text{ erg}$  and  $1 \text{ MeV} = 10^6 \text{ eV}$ ). Using these values and knowing the magnitudes of the mass defect, we can easily calculate the binding energy of an atomic nucleus.

Figure 245 shows a curve of binding energy per nuclear particle, i.e.  $c^2 \Delta m/M$ , as a function of mass number. It is seen that the binding energy per nuclear particle first rises rapidly, although not quite uniformly, then remains at approximately 8 MeV and finally drops slightly for the last elements in the Mendeleev periodic table. The following conclusion may be drawn from the fact that the energy remains constant at 8 MeV over a large portion of the curve: since the binding energy per particle is independent of the total number of particles in a nucleus, interaction in a nucleus must occur only between particles that are very close to each other.

Hence it follows that nuclear forces between particles become effective only when one particle approaches another very closely (see below).

It is instructive to compare the 8 MeV value with the chemical binding energies of molecules. The latter are usually equal to several electron volts per atom. Hence, the energy required to break up a nucleus is several million times greater than the energy required to break up a molecule into atoms.

Nuclear forces will be considered in greater detail below. It is already clear from the above examples, however, that these forces reach tremendous values when a nucleus breaks up. It is also obvious that nuclear forces constitute a new kind of force, since they can bind together particles the electric charges of which are of the same sign. Nuclear forces are not reducible to electrical forces.

#### Sec. 214. SPIN AND MAGNETIC MOMENT OF A NUCLEUS

Nucleons, the components of a nucleus, have a spin and, hence, a magnetic moment. Thus, the presence of spin is not peculiar to an electron alone. Elementary particles may have spin, and a visual interpretation of this fact is not only unnecessary, but incorrect. We have already indicated that the model of a particle rotating about its axis is completely without basis since the spin of a particle cannot be interpreted classically.

The angular momentum of any particle and, hence, of an atomic nucleus can be represented by the formula

$$\sqrt{s(s+1)} \times \frac{h}{2\pi},$$

and the projection of the spin on the selected direction line may attain  $2s+1$  values within the interval of  $s$  to  $-s$ . Usually, it is not the above expression which is called spin, but the number  $s$  determining this expression.

In accordance with the laws of quantum mechanics, the difference  $2s$  between the largest and smallest values of spin must equal a whole number or zero. Therefore the spin of a particle may equal  $0, \frac{1}{2}, 1, \frac{3}{2}, \dots$ .

Neutrons and protons, just like electrons, have a spin of  $\frac{1}{2}$ .

Examining tables of spin values for various atomic nuclei, we observe a number of interesting regularities. First of all, nuclei consisting of an even number of protons and an even number of neutrons (for instance, He, C<sup>12</sup>, O<sup>16</sup>) have zero spin. Evidently, the number of nucleons equal to a multiple of four generally plays an important role. In many cases, but by no means in all, the spin of an atomic nucleus can be determined in the following manner: the number closest to  $M$  which yields a whole number when divided by four is subtracted from  $M$  and the result is multiplied by  $\frac{1}{2}$ . For example, Li<sup>6</sup> has a spin equal to  $2 \times \frac{1}{2} = 1$ , Li<sup>7</sup> —  $\frac{3}{2}$ , B<sup>10</sup> — 1, and B<sup>11</sup> —  $\frac{3}{2}$ .

There are no exceptions to the following, rather obvious, rule: the spin of a nucleus for which  $M$  is even is equal to a whole number or zero, and the spin of a nucleus for which  $M$  is odd is equal to an odd multiple of  $\frac{1}{2}$ .

The spin of an atomic nucleus is determined from the hyperfine structure of its optical spectrum. Even though the energy levels split to a very small extent, the differences in levels may be measured very accurately. Splitting occurs due to the fact that different mutual orientations of electron spin and nuclear spin correspond to different energies.

Available data on nuclear spins indicate that the Pauli exclusion principle is applicable to the protons and neutrons of a nucleus. Two identical particles can be located at a single energy level only if their spins are anti-parallel. Since a proton differs from a neutron, two protons and two neutrons may be located at a single level. This compact group of four particles, having a total spin of zero, will be recognised as the nucleus of a helium atom ( $\alpha$ -particle).

If a particle has a spin, it also has a magnetic moment. The angular momentum  $L$  must be directly proportional to the magnetic moment  $M$ , but the magnetic moment may be either parallel or anti-parallel to the spin.

If the spin of a particle (simple or complex) is  $s$ , then its magnetic moment may be written in the form

$$M = g\mu s,$$

where  $\mu$  is an elementary magneton equal to  $\frac{eh}{4\pi mc}$ ,  $m$  is the mass of the particle, and  $g$  is a dimensionless factor. This equation is the generalised form of the relation given in Sec. 194 for the case of an electron. For such a particle  $s = \frac{1}{2}$  and  $g$  must be set equal to 2 in order to obtain agreement with available experimental data.

Different particles (elementary as well as complex) have different values of  $g$ . For instance, the  $g$ -factor of a neutron is equal to 3.8206, and that of a proton is equal to 5.5791.

The value of an elementary magneton depends on the mass of the particle. However, it is customary to use only two magneton values: the Bohr magneton for light particles and the nuclear magneton (calculated for the case of a proton) for heavy particles, whereby  $\mu_N = \frac{1}{1,836} \mu_{\text{Bohr}}$ . The above indicated values of the  $g$ -factor are calculated for  $\mu_N$ .

A theory of  $g$ -factors and magnetic moments which relates these properties of a nucleus to its structure does not exist.

## Sec. 215. NUCLEON INTERACTION FORCES

Our basic knowledge of nuclear forces can be obtained by studying the scattering of particles. It was concluded from Rutherford's first experiments in the scattering of  $\alpha$ -particles that nuclear forces have a very small range. Rutherford's results could be explained quantitatively by assuming that the deflection of  $\alpha$ -particles is the result of electric repulsion between charged particles having the same sign. The experimental results agreed with the results of theoretical calculations even when an  $\alpha$ -particle passed extremely closely to the scattering nucleus. This means that it suffices to separate two nuclear particles by a very small distance in order for the effective forces to consist only of electric forces; the nuclear forces will then be no longer effective.

More direct evidence can be obtained from the scattering of neutrons by protons. For this purpose a neutron beam is passed through gaseous hydrogen. Experiments show that only a small fraction of the neutrons collide with nuclei of hydrogen atoms. The angular distribution of scattered neutrons is uniform. This result differs basically from that of  $\alpha$ -particle scattering, i.e. scattering due to electric interactions. In such scattering, deflection always occurs, but the deflection is small when an  $\alpha$ -particle passes far from a nucleus and large when such a particle passes close to a nucleus. It can be concluded from the patterns obtained in the scattering of neutrons by protons that the effective range of nuclear forces is very



small. A value of the order of  $2 \times 10^{-13}$  cm is reliably deduced from such experiments.

This is also the value obtained for the effective range from the scattering of protons by protons. In this case the experiments and calculations are rather intricate, since it is necessary to "deduct" that portion of the scattering which is due to purely electric interaction. However, the deduction required can be determined using data from observations at high energies and large angles. Unfortunately, direct experiments in the scattering of neutrons by neutrons are not possible, but considerable indirect evidence indicates that in this case too nuclear forces have the same properties. For example, let us compare the binding energies of tritium (hydrogen isotope of mass 3) and the helium isotope of equal mass. In the first case a nucleus consists of two neutrons and one proton and in the second of two protons and one neutron. It turns out that the binding energy of this helium isotope is higher than that of tritium by an amount exactly equal to the electric interaction between two protons.

All these experiments and reasonings lead to a conclusion that the nuclear forces acting between nucleons are independent of the electric charges of the interacting particles.

From experiments in the scattering of nucleons one can also conclude that the interaction is of an exchange nature. The term 'exchange' is used to indicate that the colliding particles interchange properties, i.e., a proton is transformed into a neutron and vice versa. Experimentally, this has been shown to occur in the scattering by protons of a beam of neutrons having very high energies (scores of times greater than the potential energy of interaction between protons and neutrons). One would expect that in such an experiment most neutrons pass through hydrogen without scattering, but in fact the forward scattered beam consists of equal numbers of neutrons and protons.

The problem of nuclear forces is considerably complicated by the fact that such forces depend on nuclear spin orientations. The latter cause nuclear forces to lose their central nature, i.e., the forces do not act along the lines joining the centres of particles.

Nevertheless, by averaging this dependence, we may characterise nuclear forces by interaction potential of the same type as was discussed for molecules. Figure 246 shows such a curve for two nucleons, the interaction energies are plotted in MeV and the distances in fermis ( $1 \text{ fermi} = 10^{-13} \text{ cm}$ ). At a distance of 4 fermis the nuclear forces cease their action. A dashed line in the graph is given to compare nuclear interaction with electrostatic interaction which represents an electrostatic potential of two opposite charges each equal to 3.3 charges of an electron. At a distance of 2 fermis the interaction energies are equal to each other. Certainly, there is no analogy between the two curves.

Such curves are plotted from experimental data obtained for different nuclei. The ordinates and abscissas of potential wells vary within small ranges (5 to 20 eV, several fermis).

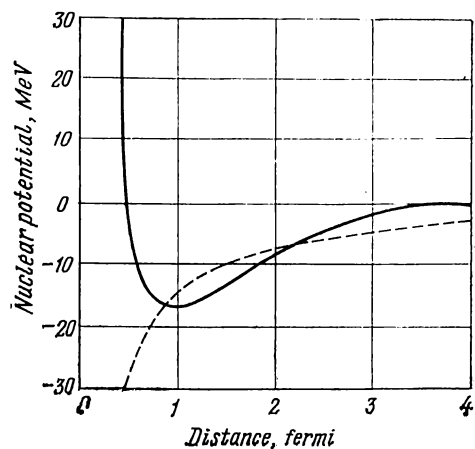


Fig. 246

## Sec. 216. NUCLEONS IN A NUCLEUS

There is no doubt that nucleons are very closely spaced in a nucleus. In reasonably close agreement with a number of experimental findings is the following formula for the "radius" of a nucleus:

$$R = k^3 \sqrt[3]{M},$$

where  $M$  is the mass number, and  $k = 1.5 \times 10^{-13}$  cm. The formula has been derived for heavy nuclei, but it is undoubtedly valid for light nuclei as well. Since the nuclear radius is proportional to the cube root of the number of particles in a nucleus, the volume is proportional to the first power of this number. Therefore it must be assumed that, at least to a rough approximation, the packing density of nucleons in a nucleus is uniform.

Since nucleons are of a wave nature, one cannot, of course, construct a geometric model of an atomic nucleus and determine the path of nucleons in a nucleus.

As an approximation, each nucleon can be pictured as moving in the field of all the others. Such a nucleon will have a system of energy levels which can be filled consecutively in going from light nuclei to heavier ones. Like in the case of electrons, the lowest level cannot have an angular momentum. In accordance with the Pauli exclusion principle, at this lowest level we can have two neutrons and two protons (the particles of each pair having opposite spins), i.e., an  $\alpha$ -particle. An analogous analysis of heavier nuclei shows that stable groups of particles will also be located at other levels. The shell model of a nucleus has proved very useful in determining a number of nuclear properties and in explaining the prevalence of various isotopes.

In examining the composition of atomic nuclei, as we proceed from light particles to heavier ones, we pay attention to the following fact: the number of neutrons in an atomic nucleus increases more rapidly than the number of protons. For instance, the nucleus of the lead atom (which is a very stable element) contains 82 protons and 126 neutrons. This increase in the number of neutrons is explained by the need to counterbalance the increased electric repulsion of the protons.

On the other hand, the lack of equality between protons and neutrons causes a disadvantageous effect. Indeed, when this equality existed, the low-energy levels could be filled with a maximum number of particles, since, according to the Pauli exclusion principle, two neutrons and two protons can exist in one state. However, if this occurred, the electric repulsion would increase too much and the total energy would not be a minimum. Evidently, an actual case represents a compromise solution between these two tendencies. Beta-decay phenomena, which occur so frequently in radioactive elements, represent the selection of the optimum situation in the indicated sense. If there are too many protons or neutrons in a nucleus, the situation is corrected by the emission of an electron or a positron.

## Sec. 217. SPECTRA OF ATOMIC NUCLEI

From the viewpoint of quantum mechanics, any assembly of interacting particles can be characterised by a 'hierarchy' of energy levels. An atomic nucleus may occupy the lowest level (as it is the case with all stable elements under ordinary terrestrial conditions), but it may also jump to an excited level on undergoing collisions with other particles. A nucleus can reside in an excited energy level only for a short period of time, then it jumps to a lower energy state, which is accompanied by a release of extra energy.

The nuclei of radioactive atoms are usually either in an excited state, or jump into an excited state on emitting an  $\alpha$ -particle (see below).

The energy differences between the levels in nuclei are of the order of hundreds of thousand electron volts (as compared to several electron volts in the case of atoms). This is why, under ordinary conditions, nuclei are quite inactive. Energy transitions occurring in nuclei are studied in detail by bombarding them in accelerators designed to operate at millions of electron volts.

Figure 247 shows energy level systems of two nuclei:  ${}^5\text{B}^{11}$  and  ${}^6\text{C}^{11}$ . The energy levels are sharply defined and measured with a high accuracy. Energy values are given in the figure.

A pair of nuclei with equal total number of nucleons and in which the number of neutrons (protons) in one nucleus is equal to the number of protons (neutrons)

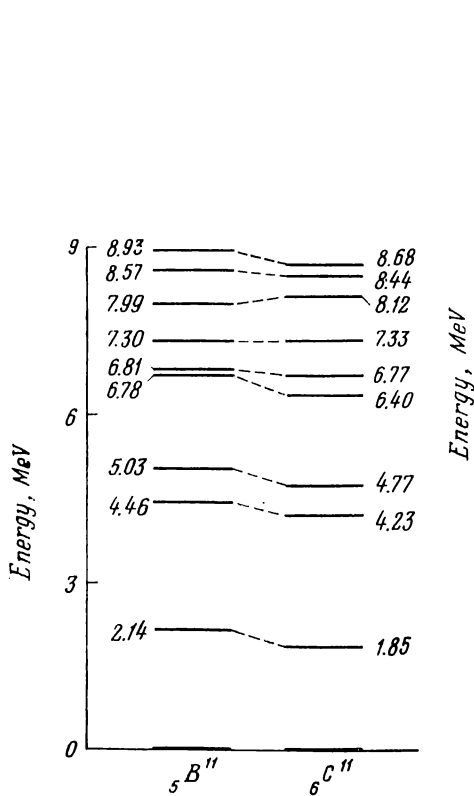


Fig. 247

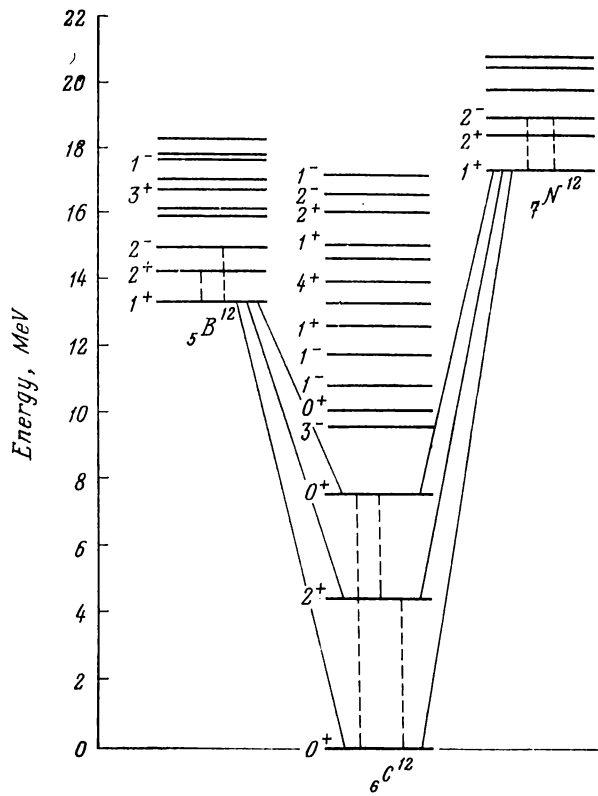


Fig. 248

in the other is called a "mirror" pair. Obviously, the difference in nuclei is reduced to electrostatic energy which is not high. The level patterns turn out to be very close to each other; this once again illustrates the generality of nucleon interaction.

At each excited state a nucleus possesses a certain combination of properties. To different energy levels there correspond, generally speaking, different spins, parities, and isospins. *Isospins* are defined as numbers which number the levels in each isospin multiplet. This quantum number will be discussed below when we consider nucleon spectra. *Parity* is a particle property associated with its

helical symmetry (see Sec. 234). Two variants (a left-handed screw and a right-handed screw) are designated by the signs  $+$ , and  $-$ , respectively.

Passing over to a lower level, an atomic nucleus may release its energy in two different ways: (1) by emitting a photon, i.e., in the same way which is the only possible in the case of atom radiation. Since the energy difference  $E_2 - E_1$  is very great, there appear "hard" photons which are usually called gamma rays; (2) by emitting a so-called pair of leptons. This pair usually consists either of an electron and an antineutrino, or a positron and a neutrino (cf. Sec. 233).

A level transition pattern for a nucleus consisting of twelve nucleons is shown in Fig. 248, where the numbers indicate spins of states, and the signs (plus or minus) denote parity. The figure represents energy levels of three nuclei. These three nuclei should not be considered separately, but just as a single system with all possible transitions. The vertical (broken) lines show gamma transitions; the horizontal continuous lines indicate the transitions accompanied by emitting a pair of leptons. The lines drawn from left to right downward correspond to emission of an electron and an antineutrino and those drawn from right to left downward to emission of a positron and neutrino. According to the law of conservation of a charge, during the emission of a pair of leptons the proton is converted into a neutron, and vice versa.

In the case of omission of a pair of leptons, the difference between energy levels must be not less than the energy corresponding to the electron mass, i.e., not less than 0.51 MeV (the rest mass of neutrino is equal to zero). The excess energy (if any) will be converted into kinetic energy of the leptons.

#### Sec. 218. NEUTRINO EMITTED IN BETA-DECAY

An energy transition accompanied by emission of a lepton pair is called beta-decay. Neutrino detection is a complex experimental problem solved only during the last two decades. Originally, transitions similar to those shown in Fig. 248, were described as a transformation (decay) of one (radioactive) nucleus into another (stable) nucleus with an electron emitted. But long before neutrino was discovered, the necessity of its existence had already become obvious.

It was impossible to doubt the validity of the law of conservation of angular momentum, as well as of the law of conservation of energy. Indeed, the neutron, proton, electron, and positron—all these particles have a spin of  $\frac{1}{2}$ . As has been indicated,  $\beta$ -decay in an atomic nucleus transforms a proton into a neutron, or vice versa. Since the number of nucleons in a nucleus remains unchanged  $\beta$ -decay cannot transform an even spin into an odd one. But this is precisely what would be required if only an electron, having a spin of  $\frac{1}{2}$ , were ejected during  $\beta$ -decay. This contradiction was resolved by hypothesising the existence of the neutrino, a particle having a spin of  $\frac{1}{2}$ . In addition, by means of the neutrino, we can explain why the  $\beta$ -particle spectrum is continuous.

If  $\beta$ -decay consisted in the ejection of only an electron, then this electron would have a well-defined energy, since the initial and final energy levels, i.e., the energies of the primary nucleus and the new nucleus, are well-defined. But, as has been indicated, a continuous spectrum of electrons, from a certain maximum velocity down to zero, is obtained. Such a spectrum can be explained by assuming that during disintegration two particles are ejected from a nucleus according

to the equation  $n \rightleftharpoons p + e + \bar{\nu}$ . The energy is divided between the electron and the neutrino in a random manner.

As has been mentioned, the detection of a neutral particle having a negligibly small mass (we now know that the mass of a neutrino is less than 0.002 of the mass of an electron) is an extremely difficult problem. This problem was not solved until 1956.

It should be noted here that the particle which had been given the name 'neutrino', later on, upon the discovery of the law of antisymmetry of elementary particles, turned out to be an antineutrino and was denoted as  $\bar{\nu}$  (see Secs. 233 and 234).

We have indicated that a neutrino is formed when a neutron disintegrates. A particularly large number of such decay events should occur in a nuclear reactor, where a tremendous number of nuclear fragments, rich in neutrons, are continuously formed. If antineutrinos exist, then a stream of such particles should emerge from a reactor. When a neutrino  $\bar{\nu}$  collides with a proton, the following reaction should occur:  $\bar{\nu} + p \rightarrow e^+ + n$ , i.e. a positron and a neutron will be formed. Such reactions should be observed in targets containing large numbers of hydrogen atoms (that is, protons) if they are placed close to a nuclear reactor. This reaction should occur very rarely (several times per hour), since a neutrino has extremely high penetrating power. At the same time a large number of other nuclear reactions occur close to a reactor.

The difficulties involved in detecting a neutrino are evident. They can be overcome by properly utilising the distinguishing features of this reaction. We know that the positron is quickly annihilated with an electron of one of the target atoms and that such annihilation yields two photons. The neutron, after covering a certain distance in the target, is absorbed by one of the impurity atoms (cadmium) added to the target for this purpose. The average life-time of a neutron before being absorbed has been calculated. It is equal to approximately five microseconds. The absorption of a neutron by cadmium is accompanied by gamma radiation. Using modern measuring techniques, the experimenter must distinguish the following sequence of events from all others: the simultaneous creation of two photons, followed in five microseconds by a stronger pulse of gamma radiation. Since this has been achieved, the existence of the neutrino may be considered to be a fact.

#### Sec. 219. GENERAL LAWS OF CHEMICAL AND NUCLEAR TRANSFORMATIONS

We shall now discuss several general energy relationships which are equally applicable to chemical reactions and to transformations of atomic nuclei and other particles.

Transformations can occur only when particles closely approach each other. Since particles must possess a certain kinetic energy in order for a transformation to occur, we are quite justified in using the term "collision" to describe a close approach between particles. Not every encounter between particles results in a transformation. The mechanism of chemical and nuclear transformations is very difficult to study. Since direct observations are impossible, one is forced to make hypothesis the validity of which can be checked indirectly.

In the case of chemical transformations, the mutual orientation of molecules upon collision undoubtedly plays an important role. For a certain reaction to occur, molecules must approach each other in a manner suitable for the regrouping of atoms.

For every transformation which can occur on a mass scale (the usual case in chemical and nuclear reactions where billions of molecules or nuclei collide during a short interval of time) one can indicate, in principle, the number  $A$  which, generally speaking, will be a measure of the fraction of encounters in which particles are in a position "suitable" for a transformation to occur.

However, the requirement of appropriate orientation is, of course, not the only condition to be met for a transformation to occur. Since a particle is ordinarily stable and, hence, possesses a minimum of potential energy, an energy sufficient to lift the molecule out of its potential well must be imparted to it. This minimum energy is called *the activation energy*. Figure 249 shows a potential energy curve. The particle is stable when  $r = 0$ , where  $r$  is a fixed parameter. In order for a reaction to occur, an activation energy  $\mathcal{E}$  must be provided. For the case represented in the figure the reaction proceeds with a liberation of heat.

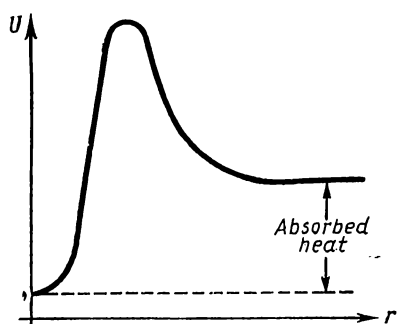


Fig. 249

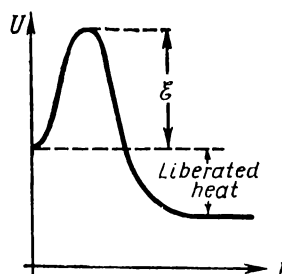


Fig. 250

We may apply the Boltzmann law to collisions between molecules or between nuclei (under similar conditions), i.e., we can assume that the number of encounters resulting in a transformation is proportional to  $e^{-\mathcal{E}/kT}$ , where  $\mathcal{E}$  is the activation energy.

Obviously, the rate of transformation may be expressed as the product  $Ae^{-\mathcal{E}/kT}$ , where the first factor takes into account the "geometric" conditions of encounter and the second the energy aspect. It is customary to give special consideration to the case of two colliding particles at the instant when the potential energy is maximum. Such an activated complex (as it is referred to in chemistry) or compound nucleus (as it is referred to in the study of nuclear transformations) exists but for short instants. The obtained system can "slip back" into the potential well or "roll over" the well wall. In the latter case, a transformation has occurred and a new system with a new potential energy has been formed.

In chemical as well as nuclear transformations, the resulting system may consist of a single particle (addition reaction) or two new particles.

If the potential energy of the new particles is greater than the potential energy of the original particles (cf. the bottom of a volcano crater is below the level of the foot of the mountain on which the volcano is located), the transformation proceeds with the absorption of energy. The absorbed energy (heat) will be equal to the difference between the activation energy and the energy of the reaction products (see Fig. 250). If the energy of the created particles is less than the energy of the original particles, heat is liberated.

Both kinds of transformations are encountered in chemistry and in nuclear physics. Reactions proceeding with liberation of heat are called *exothermic*, while those proceeding with absorption of heat are called *endothermic*.

Chemical and nuclear transformations are often accompanied by radiation. However, as a rule, the main energy effect of a reaction consists in the transformation of the potential energy of the atoms in a molecule (or nucleons in a nucleus) into kinetic energy of particles. Therefore, generally speaking, transformations in which heat is liberated are those in which two slow particles collide and produce two fast particles. Of course, in endothermic reactions, the opposite takes place.

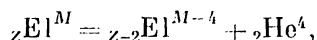
It is seen from the formula that the rate of transformation increases exponentially with temperature. This is the reason why chemical transformations are extremely sensitive to changes in temperature. The higher the temperature, the stronger the impacts of colliding particles. It is well known that temperature plays an important role in chemical transformations. In nuclear transformations, as a result of the tremendous values of binding energy, the role of temperature change is not so noticeable. The activation energy of atomic nuclei has an order of magnitude of several MeV, but if, for example, the temperature is increased by 3,000°C the energy of an atomic nucleus increases by only 0.4 eV.

The temperature must be increased to millions of degrees rather than to thousands if we wish to accelerate nuclear transformations (see below).

## Sec. 220. RADIOACTIVITY

Radioactive disintegration is the simplest nuclear reaction. It consists in the ejection of an  $\alpha$ -particle from an atomic nucleus.

Since an  $\alpha$ -particle consists of two protons and two neutrons, its symbol is  ${}_2\text{He}^4$ . Thus, alpha decay may be represented as follows:

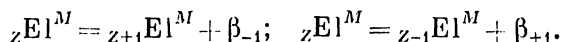


where El is an arbitrary chemical element. This disintegration reaction, as well as other nuclear reactions to be considered in the following section, obeys the law of conservation of charge (the sum of the subscripts in the right member of the above equation must be equal to  $Z$ ) and the law of conservation of mass number (the sum of the superscripts in the right member must equal  $M$ ).

Both an initial and a final nuclei may be in an excited state. Therefore, alpha decay may be accompanied by energy transitions either with emission of photons, or with formation of lepton pairs.

Owing to historical reasons, beta decay is rightfully considered parallel with other nuclear reactions. Neglecting for a while neutrino (whose existence is of no importance for a chemist), we may assert that beta decay consists in the ejection of an ordinary electron ( $\beta^-$ ) or a positron ( $\beta^+$ ) from a nucleus. It should be recalled that the masses as well as the magnitudes of the charges of these two particles are equal. The ejection of such a light electrically charged particle from a nucleus results in the nuclear transformation of a proton into a neutron or a neutron into a proton. Such a transformation ensures the conservation of electric charge upon disintegration.

Beta decay patterns are written in the following form:



Thus, for  $\beta^-$ -decay, a neutron of the nucleus is transformed into a proton (the number of protons increases), for  $\beta^+$ -decay, the reverse transformation takes place.

In addition, gamma rays, i.e., electromagnetic radiation of shorter wavelength than X-rays, may be emitted in both types of radioactive disintegration. Alpha radioactivity occurs only for the heavy elements (beginning with bismuth);  $\beta^-$ -radioactivity is encountered considerably more often than  $\beta^+$ -radioactivity.

A radioactive substance found in nature is said to be *naturally radioactive*; a radioactive substance obtained by means of nuclear reactions is said to be *artificially radioactive*.

If another element is formed when the nucleus of a radioactive element decays, and if a third element is formed from the second, etc., the sequence of such elements is called a *radioactive series*. Four radioactive series, beginning with  $U^{238}$ ,  $Th^{232}$ ,  $U^{235}$ , and  $U^{233*}$ , are known.

Radioactive decay obeys the law

$$N = N_0 e^{-\lambda t},$$

where  $N_0$  is the number of nuclei present at the initial instant  $t = 0$ ,  $N$  the number of nuclei remaining (not disintegrated) after the elapse of a time  $t$ , and  $\lambda$  a *radioactive-decay constant* (it does not change for a given element).

It is easily seen that the time interval  $T$  during which half of the number of atoms present decay is given by

$$T = \frac{\ln 2}{\lambda} = \frac{0.693}{\lambda}.$$

The time interval  $T$  is called the *half-life period* or simply *half-life* of a radioactive element.

The ancestors of the naturally radioactive series have a half-life lying within the range of  $10^8$  to  $10^{10}$  years. On the other hand, the half-life of intermediate decay products and artificially radioactive elements may be equal to an extremely small fraction of a second.

A mass of radioactive substance could be expressed, of course, in grams. However, it is easier and more convenient to characterise it by its activity, i.e., the number of decay events per second. The curie is a historical unit of measurement which is equal to  $3.7 \times 10^{10}$  decays per second. In laboratory work, this unit is found to be inconveniently large; hence, the millicurie (that is, one-thousandth of a curie) is frequently used. Another unit used is the rutherford, which is equal to  $10^6$  decays per second. Thus, one millicurie is equal to 37 rutherfords.

If the half-life  $T$  of a substance is known, its initial radioactivity can be easily calculated. The fraction of substance decaying in one second is equal to

$$1 - \frac{N}{N_0} = 1 - e^{-\lambda}$$

or, since  $\lambda$  is a small number, we get

$$1 - \frac{N}{N_0} \approx \lambda = \frac{0.693}{T}.$$

The activity of  $m$  grams of a substance of atomic weight  $A$  is equal to

$$\frac{0.693}{T} \cdot \frac{m}{A \times 1.66 \times 10^{-24}} \frac{\text{decays}}{\text{second}}.$$

It is generally assumed that the radioactive products found during 100 days of operation of a nuclear reactor (see below) amount to 1 curie per watt. For 500,000 kW, the amount of decay products is equal to about 500 g, i.e.,  $10^{-6}$  gram per watt. Taking the average atomic weight of the fission products as 100, we find by the above formula that the average half-life of the radioactive products is equal to  $10^5$  seconds, i.e., about 24 hours.

\*  $U^{233}$  has predecessors among the transuranic elements.



We cannot describe fully the properties of a radioactive substance by means of its half-life and activity alone. We must indicate, in addition, whether the substance is an  $\alpha$ -particle or  $\beta$ -particle radiator and whether the disintegration is accompanied by  $\gamma$ -radiation. An even fuller description requires that we specify the energy of the particles ejected from the nuclei and the energy of radiation. The properties of  $\alpha$ -particles radiated by different radioactive materials differ but very little. Their initial velocities lie within the range of 15,000 to 20,000 kilometres per second and the number of pairs of ions formed in air by such an  $\alpha$ -particle lies in the range of  $1 \times 10^5$  to  $2 \times 10^5$ . The energies of  $\beta$ -particles ejected during decay are distributed continuously from zero to several hundred or thousand keV. The energies of gamma rays emitted by one radioactive substance differ from the energies of gamma rays emitted by another radioactive substance, but their order of magnitude remains the same for all elements.

In alpha decay an  $\alpha$ -particle tunnels through a potential barrier and is then subjected to electrostatic repulsion. The potential curve of a nucleus is shown in Fig. 251. We see a potential well and a potential barrier; beyond the barrier the electrostatic potential energy drops hyperbolically. It has been shown in the case of radioactive element by scattering alpha particles from nuclei of this element that the height of the potential barrier exceeds 9 MeV. At the same time, particles having an energy of only 4 MeV escape from nuclei by tunneling through the barrier.

Such a model explains why one radioactive element has a very short half-life, while another has a very long half-life. If the difference between the energy of an  $\alpha$ -particle in a nucleus and the height of the potential barrier changes even slightly, the probability of  $\alpha$ -particle leakage through the barrier changes radically (see the formula in Sec. 188).

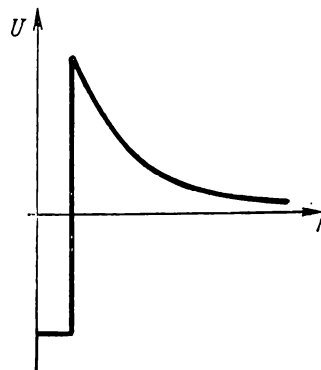


Fig. 251

Alpha and beta decay obey the following formula of decay as a function of time:

$N = N_0 e^{-\lambda t}$ . In fact, the decay of a nucleus is an independent event that does not affect the behaviour of other nuclei. All nuclei have the same decay probability. Let us assume that half of the nuclei disintegrate during a time interval  $t$ . But the remaining half is subject to the same conditions as the original group of atoms, and consequently, half of the remaining half disintegrate during an equal interval of time. The fact that the decay of a nucleus is not dependent on the behaviour of its neighbours means that in a given time interval  $\Delta t$  the fraction of the number of atoms present which disintegrate, i.e.,  $\frac{\Delta N}{N}$ , will always be the same. This statement may be written in the form

$$\frac{\Delta N}{N} = -\lambda \Delta t.$$

By integrating this expression, we obtain the exponential law of decay.

It is useful to remember that the reason why exponential laws are encountered so frequently in physics is that they are the mathematical expression for decrease in accordance with the widespread rule that for equal changes in argument a function decreases by the same fraction of its magnitude.

## Sec. 221. NUCLEAR REACTIONS

A considerable amount of experimental data has been accumulated on nuclear reactions. The number of such reactions which have been studied reaches several thousand.

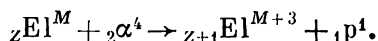
At present, the following types of nuclear reactions are known (in addition to radioactive decay, which may be considered as a nuclear decomposition reaction): capture reactions, in which two colliding particles combine; exchange reactions, in which a particle is captured and another is ejected; and fission reactions, in which a nucleus breaks up as the result of energy received in one or another form. Nuclear reactions which occur under the action of hard  $\gamma$ -rays are known as *photonuclear reactions*.

With the aid of nuclear reactions, it is possible to obtain stable natural isotopes as well as unstable radioactive isotopes which are not encountered in nature. Among the latter isotopes, it has proved possible to synthesise elements which have no stable isotopes at all (for instance, technetium, an element having the atomic number 43), and also transuranic elements.

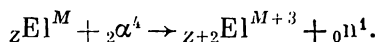
The reactions which occur when various nuclei are bombarded with  $\alpha$ -particles, protons, and neutrons have been studied most carefully.

When an  $\alpha$ -particle collides with a nucleus, one of two types of reactions generally occurs: either the  $\alpha$ -particle is captured and a proton (p) is ejected or the  $\alpha$ -particle is captured and a neutron (n) is ejected. These reactions are designated by the symbols  $(\alpha, p)$  and  $(\alpha, n)$ , respectively.

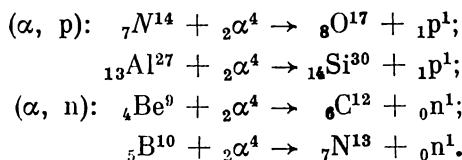
The equation for an  $(\alpha, p)$  reaction has the form



The equation for an  $(\alpha, n)$  reaction has the form

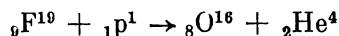


Here are some examples of  $\alpha$ -particle reactions:



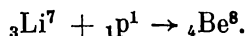
The first of the above  $(\alpha, n)$  reactions is of great practical importance, since a mixture of radium ( $\alpha$ -particle source) and beryllium is a common neutron source.

A large class of reactions occur as the result of collisions with protons. Such reactions include  $(p, \alpha)$  reactions (in which a proton is captured and an  $\alpha$ -particle is ejected), e.g.,



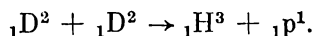
and  $(p, n)$  reactions (in which the ejected particles are neutrons).

Capture reactions which are not accompanied by the ejection of a particle also occur. The excess energy is released in the form gamma rays. Therefore, such reactions are designated as  $(p, \gamma)$  reactions, e.g.,

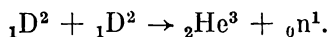


A careful study has been given to reactions involving deuterons (D), e.g.  $(D, p)$  and  $(D, n)$  reactions. When heavy hydrogen (deuterons) is bombarded with

deuterons, a radioactive isotope of hydrogen, viz., tritium, may be formed:

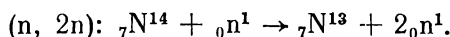
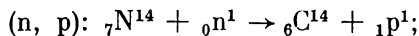


However, a (D, n) reaction may also occur:



Reactions involving neutrons are of great importance in nuclear engineering, since they occur abundantly in nuclear reactors. These include (n,  $\alpha$ ), (n, p), (n, 2n), and (n,  $\gamma$ ) reactions. In addition, reactions involving the fission of heavy nuclei occur under the action of neutrons (see below).

Two kinds of reactions occur between nitrogen and neutrons:



The first of these reactions yields a carbon isotope of long lifetime (more than 5,000 years). This isotope is of great importance in biochemical investigations.

Almost all isotopes are capable of capturing a neutron ((n,  $\gamma$ ) reaction). In this manner, an isotope of the same element, the mass of which is one unit greater than the original isotope, is formed. Usually radioactive isotopes ( $\beta$ -radioactivity) are produced.

Exothermic and endothermic reactions are possible in nuclear chemistry, like in molecular chemistry. The magnitude and sign of the thermal effect can be determined using the principle of the equivalence of mass and energy.

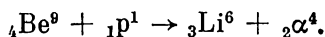
If nuclei of masses  $M_3$  and  $M_4$  are formed from nuclei of masses  $M_1$  and  $M_2$  then

$$M_1 + M_2 = M_3 + M_4 + \Delta m.$$

When  $\Delta m > 0$ , i.e., when the nuclei of the reaction products have less mass than the original nuclei, the reaction is exothermic. If  $\Delta m < 0$ , then the reaction is endothermic. The energy released or absorbed during a reaction can be determined by the formula  $\mathcal{E} = c^2\Delta m$ . Calculated and experimental results are in perfect agreement in all cases.

In most nuclear reactions, the thermal effects are of the order of millions of electron volts for each pair of reacting nuclei. This is millions of times greater than the corresponding values in chemical reactions.

Sample calculation. Let us consider the reaction



The following are tabulated handbook values of the masses occurring in this reaction:  $m_{\text{Be}} = 9.01464$ ,  $m_{\text{p}} = 1.00807$ ,  $m_{\text{Li}} = 6.01671$ , and  $m_{\alpha} = 4.00372$ . The reaction is exothermic since  $\Delta m = 0.00227$ , i.e.,  $\Delta m > 0$ . Since an atomic mass unit corresponds to 931.8 MeV, the thermal effect is equal to  $0.00227 \times 931.8 = 2.12$  MeV or about  $8 \times 10^{-14}$  calories (1 MeV =  $3.827 \times 10^{-14}$  calories).

## Sec. 222. FISSION REACTIONS OF HEAVY NUCLEI

The atom nuclei heavier than the nuclei of tin atoms are capable of fissioning into two parts of approximately equal masses. For such a fission to occur it is necessary to impart to the nucleus a considerable amount of activation energy which decreases with an increase in the mass number of the isotope. For the nuclei of uranium, thorium, and palladium this energy is close to 5 MeV. The fission reaction of these nuclei is of the exothermic character; the energy released is considerably greater than the activation energy.

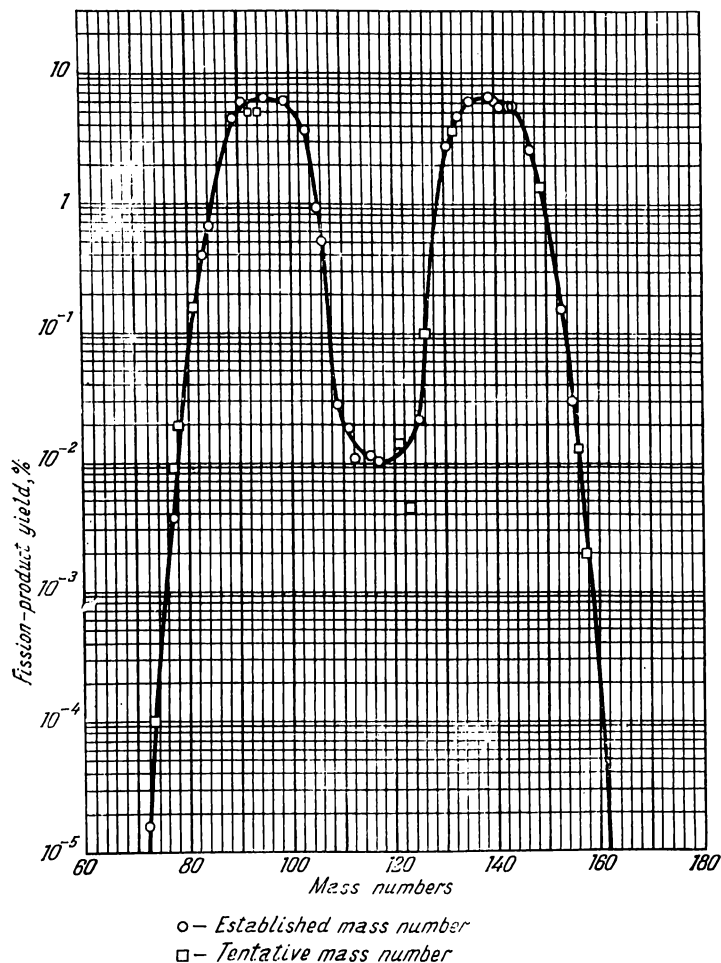
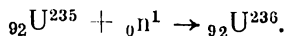


Fig. 252

The fission can be initiated by protons, deuterons,  $\alpha$ -particles, and  $\gamma$ -radiation. But of high importance for practical purposes is the fission occurring when a neutron hits a heavy nucleus.

Let us consider the fission of an isotope of  $U^{235}$  under the action of neutrons. When a neutron is captured an isotope is formed whose mass exceeds the mass of a  $U^{235}$  nucleus by unity:

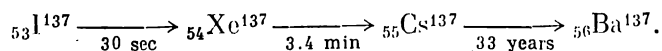


The fission will take place if the activation energy of fission of  $U^{236}$  is less than the energy delivered by the neutron. The captured neutron supplies the nucleus with energy equal to the binding energy plus the kinetic energy of its motion. In the case of  $U^{235}$  the energy delivered even by a neutron moving with a thermal velocity is sufficient for realising the fission.

High energy is required for fission of  $U^{238}$  nuclei, and the nuclei of this isotope fission only under the action of fast neutrons.

$U^{235}$  nuclei fission into two different fragments. The nucleus is fissioned in an arbitrary way, yielding various primary products. A graph of fission yield for  $U^{235}$  depending on the masses of fragments is given in Fig. 252. As is seen from the curve, fission by halves is the least probable case. Mostly, the masses of the fission fragments are related to each other as approximately two to three, say  $Sr^{95}$  and  $Xe^{139}$ . These nuclei possess huge energies: the light nucleus of the order of 100 MeV, and the heavy nucleus about 65 MeV. The diversity of fission products is explained not only by various patterns of nucleus fission, but also by the fact that the newly born radioactive nuclear fragments, in their turn, disintegrate, thus creating some more products.

The nuclei yielded by fission possess too large an amount of neutrons. Through a chain of transformations radioactive nuclei pass over to a normal stable state. For instance,



Indicated below the arrows are half-lives. It is clear, that, according to the decay time, only  $Cs^{137}$  is a practical product of this chain of transformations.

$Sr^{90}$  is another important fission product of  $U^{235}$  (see below).

Nucleus fission of  $U^{235}$  is accompanied by a release of large amounts of energy: one gram of uranium releases the same amount of energy which is obtained from the burning of 2.5 tons of coal, i.e. 22,000 kWh. The main part of energy is released in the form of the kinetic energy of fission fragments; about 10 percent of energy is obliged to radiation.

### Sec. 223. CHAIN REACTIONS

The reaction occurring in  $U^{235}$  is known in detail.  $U^{235}$  is the only natural isotope which makes it possible to use the released energy for industrial purposes.

Since the fission reaction needs neutrons, and neutron gas does not exist in nature, energy release can be realised in large amounts only under one necessary condition, i.e. in the presence of a chain reaction in the process of which new neutrons would be continuously generated. This is just the way in which  $U^{235}$  nuclei fission. During each fission several neutrons are released, their number being different. On the average one fission yields 2.5 neutrons. If the neutrons obtained as a result of fission of one nucleus are able to break the nuclei of other atoms of uranium, then a chain reaction will be realised.

Since neutrons have a considerable mean free path, the neutrons generated by a fissioned nucleus of  $U^{235}$  have a good chance to escape without breaking other nuclei. Besides, it should be taken into consideration that not every collision of a neutron with a  $U^{235}$  nucleus leads to fission.

The development of a chain reaction depends on the neutron multiplication factor\*. Theoretically, we can evaluate this factor  $K_0$  for the case of a system of infinitely large dimensions. To calculate  $K_0$  we have to know the number of the neutron multiplied due to slow and fast fission, as well as the probabilities of neutron capture by the nuclei of nonfissionable substances.

The values of the factor  $K_0$  exceeding unity indicate to what extent the next generation of neutrons are more numerous than the preceding one.

---

\* Sometimes called the effective multiplication constant and defined as the ratio of the number of neutrons present in a given stage of the reaction to that present in one stage earlier.

But a reactor has finite dimensions, therefore in this case the multiplication factor must be written in the following form

$$K = K_0 (1 - p),$$

where  $p$  is the probability of neutron escape. For a reactor to start operating it is necessary that the factor  $K$  be more than unity. In the course of operation of a reactor the multiplication factor  $K$  must be exactly equal to unity.

The dimensions of a system containing nuclear fuel are said to be critical if its multiplication factor is equal to unity.

The multiplication factor can be controlled on the basis of the following reasons. The probability of meeting another nucleus by a neutron prior to its escape can be increased by gathering large amounts of nuclear fuel. It is also necessary to reduce to minimum the number of atom nuclei capable of capturing the neutrons and, thus, of leading them out of the reaction. An increase in the probability of neutron capture can be achieved due to deceleration of the neutrons. During fission fast neutrons fly out of the nucleus, while the nucleus of uranium-235 captures best of all slow (so-called thermal) neutrons.

The peculiarity of uranium-235 consists in that its nuclei fission under the action of both fast (though with a less probability) and slow neutrons. If nuclei of a certain substance fission only under the action of fast neutrons, then the realisation of a chain reaction becomes impossible, since a neutron emitted by a fissioned nucleus and affected by an insignificant deceleration due to one-two random collisions which did not result in fission, does not participate in the chain reaction any longer.

At the present time, nuclear fuels comprise first of all such substances which enable us to realise a chain reaction with slow neutrons. To such substances there belong: the only isotope uranium-235 and two artificial elements, namely, plutonium-239 obtained from uranium-238, and uranium-233 obtained from thorium-232.

For a chain reaction to begin, it is necessary only to bring together an amount of nuclear fuel exceeding a certain minimum into a bounded volume. There is no necessity to take care of the first (starting) neutron, since, due to cosmic radiation, a small amount of neutrons are always present in atmosphere. Besides, one should not forget about the phenomenon of so-called spontaneous (i.e. occurring under the action of internal forces) fission discovered by the Soviet researchers G. N. Flerov and K. A. Petrzhak in 1939\*. It was found that, occasionally, the fission of a nucleus of uranium-235 can happen without the capture of a neutron. Lastly, a radium-beryllium mixture can also serve as a source of initial neutrons.

Observations have brought out a marked difference between the fission of uranium-238 and uranium-235. The fission of uranium-238 is induced by neutrons with a kinetic energy of at least 1 MeV, while the fission in the nuclei of uranium-235 is induced by the capture of thermal (that is, slowest) neutrons. Why this happens so can be explained as follows. The compound nucleus of uranium-239 formed by the capture of a neutron by the nucleus of uranium-238 has the ratio  $Z^2/A = 35.46$  and a fission threshold  $\mathcal{E}_f = 7.0$  MeV. The binding energy of the neutron captured by the nucleus of uranium-238 is about 6 MeV. Thus, fission in the nuclei of uranium-238 can be induced by neutrons with a kinetic energy of at least 1 MeV. For a compound nucleus of uranium-236 formed by the capture of

---

\* Using a very sensitive technique and an ionisation chamber, they were able to register pulses produced by the fission fragments of uranium not bombarded by fission-inducing neutrons.

According to Flerov and Petrzhak, the half-life of spontaneous fission should be  $10^{16}$  or  $10^{17}$  years. As will be recalled, for the spontaneous alpha-decay of uranium-238 the half-life is  $10^9$  years, or by seven orders shorter.

a neutron by the nucleus of uranium-235, the fission parameter and the fission threshold are 35.9 and 6.6 MeV, respectively. From these values we may conclude that the conditions for neutrons to induce fission in uranium-235 are more favourable than in uranium-238. Besides, the excitation energy imparted to the nucleus of uranium-235 by the capture of a neutron is about 6.8 MeV.

Thermal neutrons induce fission in uranium-233 and plutonium-239, a trans-uranic element.

## Sec. 224. NUCLEAR REACTORS

If inside a certain volume of nuclear fuel there began a chain reaction and if it cannot be controlled, then it will result in an explosion, since the number of neutrons, and together with it the amount of energy released will increase with each instance. The amount of energy released during the smallest fractions of a second will be so huge that an explosion will occur.

To release energy in constant or, at all events, controlled amounts, we have to design such an installation which would enable us to control the neutron multiplication factor. These installations are called *nuclear reactors* (formally called nuclear or atomic piles). In a nuclear reactor we must have the possibility to start a chain reaction with a multiplication factor slightly exceeding unity. Then the concentration of neutrons inside the reactor, and, hence, the power of the reactor will start growing. Reaching the desired power, we should have the possibility of making the multiplication factor equal exactly to unity. In this case the chain reaction will become self-sustaining; the number of neutrons and the energy released per unit time will remain unchanged.

Each reactor must be designed so that the neutrons obtained in it are utilised in a most effective way. But this does not mean that all of them are used for the only purpose, i.e., for nuclear fission and energy release. The substances whose nuclei are capable of absorbing the neutrons may also be introduced into the reactor. With the aid of reactions with the neutrons we can obtain a large number of necessary artificial radioactive isotopes, and, what is of high importance, artificial nuclear fuel. Thus, the nuclear reactor is an installation designed not only for producing energy, but also for obtaining artificial isotopes.

The neutrons emitted in fission have a velocity of tens of thousands kilometres per second, their thermal velocity being of the order of 1 km/s (0.025 eV). Thermal (slow) neutrons are most effective as far as fission is concerned.

The source and fissionable materials usually employed in nuclear reactors are uranium-235, plutonium-239, uranium-238, and thorium-232. The naturally occurring mixture of uranium isotopes contains 140 times more uranium-238 than uranium-235. For an insight into operation of a nuclear reactor using this natural mixture of uranium isotopes, it is important to remember the difference in the conditions under which fission can be induced in them. From the energy spectrum of the neutrons emitted in fission it has been found that their average energy is about 0.7 MeV. Such neutrons can induce fission only in uranium-235. The few neutrons whose energy exceeds the fission threshold of uranium-238 have a greater probability of inelastic scattering and their energy drops below that which is necessary to bring about fission in uranium-238. Through a series of collisions with uranium nuclei, the neutrons lose their energy in small portions, are slowed down and are either captured by the nuclei of uranium or absorbed by the nuclei of uranium-235. The absorption of neutrons by uranium-235 promotes the chain reaction, while absorption by uranium-238 switches them out of reaction, and the

chain reaction is interrupted. Calculations show that in the natural mixture of uranium isotopes the probability of the chain reaction being interrupted is greater than the probability of sustaining it. Thus, neither fast nor slow neutrons can sustain the chain reaction.

It has been found that uranium-238 possesses the property of resonance absorption of neutrons. Intense absorption, for instance, takes place at an energy of 7 eV. If a reactor is loaded with a mixture of isotopes, the slowing down of neutrons to an energy less than this magnitude becomes absolutely necessary.

From what has been said it follows that the basic elements of the design of any reactor are: the nuclear fuel, the neutron moderator, the neutron absorber (intended for the multiplication factor control), and the safety facilities to protect the attending personnel against the neutron flux and  $\gamma$ -radiation emitted during nuclear transformations occurring in reactors.

There is a considerable number of operating nuclear reactors and those in the stage of design and under construction. They may differ in many respects: (1) in the material used for fuel (pure nuclear fuel, enriched nuclear fuel, natural uranium in the form of metal or chemical compounds); (2) in the arrangement of fuel (volume lattice, rod lattice, uniform distribution of the fuel in a solution, or lime); (3) in the neutron moderator (heavy or light water, graphite, beryllium); (4) in the type of cooling (water, gas, liquid sodium, or absence of cooling). Reactors can be designed for any powers: from some fractions of a kilowatt to hundreds of thousand kilowatts. Depending on the type and amount of the neutron moderator, reactors can operate on slow (thermal) and fast neutrons.

All units of the nuclear reactor are controlled automatically. Its operation control is realised with the aid of neutron detectors built in the reactor walls. They are capable of measuring ("detecting") neutron fluxes within the range of  $1 \times 10^{10}$  to  $5 \times 10^{10}$  neutrons/(cm<sup>2</sup>  $\times$  s). Used in the detectors as a neutron-sensitive material is boron-10 or uranium-235 fused on the electrodes of the ionization chamber, or a gas—boron fluoride (BF<sub>3</sub>) filling the chamber. In the first case the chamber is filled with argon, nitrogen, helium, or air under a pressure of two atmospheres.

A very high responsibility is placed with various-type amplifiers of ionisation current in the relay mechanisms transmitting the readings from the neutron detector to the actuators which control the motion of the regulating rods, as well as the safety rods.

In the course of operation the position of the control rods must, obviously, change. The point is that as the decay products are accumulated the amount of neutron absorbing substance increases. In some cases these "poisonous" substances can be removed from the reactor automatically (for instance, in the case when these products are gases). But a gradual pulling out of the regulating rods is a necessary condition for keeping the neutron density at a constant level. In some period of time (5 to 20 months) the reactor will become so "poisoned" that its further operation will turn out to be impossible. This means that the reactor must be cleaned by removing the decay products and then reloaded with fresh fuel. To the fission products which intensely absorb neutrons there belong ruthenium-103, xenon-131 and 135, neodymium-143, samarium-149 and 151, europium-151, 152 and 155, and gadolinium-155.

The multiplication factor is obviously changed due to formation of artificial fuel.

If in addition to nuclear fuel the reactor contains a certain amount of uranium-238 or thorium, then such a reactor, along with releasing heat, produces artificial nuclear fuel: plutonium from uranium-238 and uranium-233 from thorium.



Each fission of uranium-235 in a reactor produces an average of 2.5 neutrons, one of which is needed to sustain the chain reaction. If the remaining 1.5 neutrons are absorbed by uranium-238 or thorium-232, then 1.5 plutonium-239 or uranium-233 nuclei may be produced to replace the fissioned uranium-235 nucleus. If at least one new fissionable atom is produced for each atom destroyed, the process is called *breeding*, and reactors in which the number of fissionable atoms produced exceeds that of fissionable atoms destroyed (the ratio of the two numbers is called the breeding ratio) are called *breeder reactors*.

Thermal reactors operating on uranium cannot breed nuclear fuel because out of 100 absorptions of thermal neutrons by uranium-235, fission occurs only in 84.5 cases and the maximum breeding ratio that is theoretically possible is  $2.5 \times 0.845 - 1 = 1.11$  instead of 1.5. The breeding ratio is reduced still more because the moderator absorbs some neutrons and a further number is lost by leakage from the core. In reactors with a moderator, the breeding ratio is usually less than unity. For example, it is as little as 0.32 in the first Soviet nuclear power plant.

Breeder reactors use fast neutrons and, therefore, no moderators. The core is a uranium alloy enriched with uranium-235, and a heavy metal (such as bismuth or lead) that has a low neutron-absorption coefficient. Reactor control is effected by moving the reflector or by varying the mass of the fissionable material.

Among the nuclear reactors built in the Soviet Union are fast reactors generating high-intensity neutron fluxes used for irradiation purposes, isotope production, and material testing.

As a pioneer in nuclear power generation, the Soviet Union is doing much to apply nuclear energy and nuclear reactors to other peaceful uses.

Concurrently with large-scale nuclear projects, work is under way in the Soviet Union on small-scale units. A drastic reduction in reactor size is important in portable and mobile installations, notably in those used for propulsion. Nuclear propulsion plants have already been installed in submarines and ocean-going ice-breakers.

## Sec. 225. ARTIFICIAL RADIOACTIVE PRODUCTS

A most considerable amount of radioactive products are obtained in nuclear reactors. Unfortunately, we cannot control this process with the aim to obtain desirable products, and each hour of reactor operation yields a certain amount of fission products. To such "obligatory" fission products with relatively long (sufficient for practical use) half-life there belong krypton-85 ( $\text{Kr}^{85}$ ), strontium-89 ( $\text{Sr}^{89}$ ), strontium-90 ( $\text{Sr}^{90}$ ), iodine-129 ( $\text{I}^{129}$ ), iodine-131 ( $\text{I}^{131}$ ), xenon-133 ( $\text{Xe}^{133}$ ), caesium-137 ( $\text{Cs}^{137}$ ), barium-140 ( $\text{Ba}^{140}$ ), and some others.

At the present time radioactive fission products are used for carrying out scientific researches, in various fields of engineering for product processes control and quality inspection, in medicine for medical treatment, and for some other purposes. For instance, radiography is an interesting field of application of radioactive isotopes. In flaw detection of metals the X-ray technique is replaced by cobalt-60 and caesium-137.

There are two possibilities to obtain a certain radioisotope: (1) in a nuclear reactor, (2) by means of nuclear bombardment with the aid of a particle accelerator.

Any substance placed in a reactor is acted upon by neutrons. Neutron-induced reactions which are realised most readily enable us to obtain radioactive isotopes of almost all chemical elements. Therefore in isotope production the leading role is played by nuclear reactors.

In reactors the radioisotopes are formed either due to neutron capture, or as a result of knocking protons or (less often)  $\alpha$ -particles out of bombarded nuclei by neutrons.

It should be noted that a nuclear reactor cannot produce such a variety of isotopes as it is obtained, for instance, in a cyclotron. This is quite clear, since in reactors the conditions of bombardment of atom nuclei are limited by the kind of a particle (neutron) and the range of energies. But even for producing large amounts of, say, isotope  $C^{14}$ , the use of nuclear reactors entails some drawbacks. The initial material utilised for obtaining  $C^{14}$  absorbs neutrons rather intensively, and that is why the reactor should be loaded only with limited amounts of this material. Thus, particle accelerators occupy quite an independent place in radioisotope production.

Capacity of an accelerator is characterised by the energy and amount of nuclei thrown out of it per unit time. This number can be easily calculated using the characteristics for an average ion current whose intensity is equal to  $10^{-11}$  A. Knowing the ion charge, we find easily that a cyclotron produces  $10^8$  nuclei per second, i.e., for deuterium this means  $2 \times 10^{-16}$  g.

To calculate the rate at which a certain radioactive substance will be formed, we have also to know the "effective cross-section" of the reaction. This quantity (denoted by  $\sigma$ ) is measured in square centimetres and has the following meaning. If  $S$  is the area of the bombarded sample, then  $\sigma/S$  is the probability of a target nucleus being hit by a projectile nucleus, as well as of an appropriate reaction being realised. The values of  $\sigma$  are usually close (by the order of the magnitude) to  $10^{-24}$  cm<sup>2</sup>.

Sharply increased absorption takes place in some nuclei at certain velocities of neutrons. For instance, cadmium intensively absorbs slow neutrons of energy 0.18 eV, the effective cross-section of this reaction being of the order of  $7000 \times 10^{-24}$  cm<sup>2</sup>.

The cross-section of a reaction substantially depends on the energy of a projectile particle. If the curve of dependence of the cross-section on the energy of a projectile particle has a sharp peak, then we speak of a resonance cross-section.

Radioactive elements obtained both in reactors and cyclotrons are widely applied as tracers (so-called labeled atoms) almost in all branches of science and technology.

The following isotopes are used most frequently. Cobalt-60 is a  $\beta^-$ -radioactive isotope whose half-life is 5.2 years. Its intensive  $\gamma$ -radiation is widely applied for the purposes of radioscopy and irradiation. Carbon-14 which is radioactive and whose half-life amounts to 6360 years is another isotope frequently used in biochemistry, geochemistry, as also for the study of kinetics of chemical reactions. The half-lives of phosphorus-32 and sulphur-35 are 14.3 days and 87 days, respectively. Both isotopes are  $\beta^-$ -radioactive and are used to the best effect in agriculture for studying the problems of assimilation of fertilizers by various plants.

#### Sec. 226. THERMONUCLEAR REACTIONS

Theoretical calculations indicate that atomic nuclei of almost all elements can serve, in principle, as sources of energy. It turns out that any nucleus heavier than the nucleus of a silver atom possesses more energy than the components into which it may be divided. All heavy nuclei release energy when they split up. The heavier the nucleus, the greater the magnitude of this energy. That is why uranium is the most "successful" nuclear fuel.

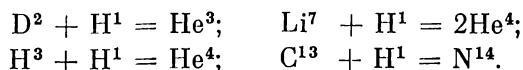
However, light nuclei can also serve as sources of energy. Theoretical calculations indicate that a nucleus obtained by the fusion of two light nuclei will possess less energy than the original particles. Therefore, energy is released upon fusion of light nuclei. Here, too, the further away from the midpoint an element stands in the Mendeleyev periodic table, the greater the amount of energy released in such a reaction. The greatest amount of energy is obtained from the fusion of nuclei of hydrogen atoms.

What conditions are required to achieve fusion between light particles? Nuclear bombardment cannot yield the desired result, since a charged particle is decelerated rapidly in a substance. The only way to achieve fusion is to raise the temperature. It is easy to calculate the temperatures that atomic nuclei require to enable them to approach each other closely, i.e., to overcome electric repulsion.

According to astronomers' calculations, the temperature at the centre of the sun is 20 million degrees. Knowing that the energy associated with each degree of freedom is equal to  $\frac{1}{2}kT$ , we can determine the average kinetic energy of a particle at this fantastically high temperature. This energy is equal to merely 3,000 eV. Now, let us calculate from the formula for the potential energy of electrical interaction,  $U = \frac{q^2}{r}$ , how closely two protons will approach each other. It turns out that the distance will be equal to  $5 \times 10^{-11}$  cm. As we know, the radius of a nucleus is considerably less than this value. Nevertheless, thermonuclear reactions, i.e., reactions occurring at a high temperature, are possible in the Sun. Calculations which take into account the tunnel effect, and the fact that in every gas, including a gas of nuclear particles, there are particles whose velocities considerably exceed the average, indicate that in a year one atom in a million takes part in nuclear fusion. This small fraction is sufficient to account for solar activity.

Under terrestrial conditions, such high temperatures have been repeatedly created during hydrogen bomb tests. Temperatures of scores of millions of degrees are created during uranium bomb explosions. If a substance whose nuclei are capable of combining and releasing energy is located within the zone of this explosion, a thermonuclear reaction, the energy of which is many times greater than that of a uranium bomb, will occur. The uranium bomb in this case serves to trigger the thermonuclear reaction.

Here are examples of the most easily realised reactions which release large quantities of energy:



Thermonuclear reactions can occur only at temperatures which give nuclei a thermal velocity sufficient to overcome, with appreciable probability, the Coulomb potential barrier.

Of greatest interest are the reactions in deuterium and in mixtures of deuterium and tritium, since these require the least amount of energy. A temperature of 200,000°C is required to obtain one neutron per second in a gram of deuterium (in accordance with the reaction  ${}_1\text{D}^2 + {}_1\text{D}^2 = {}_2\text{He}^3 + {}_0\text{n}^1$ ). In a highly rarefied gas, an even higher temperature (of the order of 500,000°C) is required for the same purpose. At such a temperature, deuterium (like other substances) will form a plasma of nuclei and electrons. To transform deuterium to this state requires very little energy—of the order of several kilowatt-hours. However, the difficulty does not lie in transmitting this energy to deuterium, but rather in providing thermal confinement, i.e., the deuterium must preserve the corresponding kinetic energy for a long period of time (see Sec. 182).

# Elementary Particles

## Sec. 227. ABOUT THE TERM "ELEMENTARY PARTICLE"

Structure of bodies is no longer a problem. All bodies consist of electrons and nuclei; all nuclei consist of protons and neutrons. Hence, the following statement will be sufficiently trustworthy and exact: bodies are built from particles belonging to the following three types—electrons, protons, and neutrons. It seems that it would be just to retain the term "elementary" only for these particles.

Problems associated with the properties of bodies constituting inanimate and animate nature are reduced to the laws of interaction of electrons and nuclei. We know these laws of nature; all physical phenomena can be explained by them, i.e., the behaviour of bodies can be predicted. If quantitative predictions turn out to be a failure, this means only that, due to a large number of factors, a phenomenon cannot be described mathematically.

The branch of physics, which explains the properties of bodies in terms of interactions of electrons, protons, and neutrons, is distinctly separating from elementary-particle physics.

A thorough study of all kinds of particle collisions occurring in cosmic rays and accelerators has led to a discovery of about two hundred particles differing in electric charge, mass, and some other properties. But in contrast to electrons and nuclei, elementary particles of this type have a short lifetime and are able to undergo a complex chain of transformations.

It seems that it would be wrong to consider all particles from one and the same viewpoint. It is more expedient to represent the elementary particles, except for the three "bricks", as excited states of a nucleon (a common name given to a proton and neutron). For such states the term "baryon" is used. Many particles and pairs of particles are analogous to light photons; they are a material representation of the energy released when an excited system returns to the principal state. In Sec. 217 we saw that a nuclear radiation may occur not only in the form of a photon, but also as a pair 'electron-antineutrino'. We are going to illustrate below that excited nucleons realise the transition to the principal state in a greater number of ways, viz., by emitting a photon, an electron-neutrino pair, a muon-neutrino pair, a pion, and a kaon.

In our opinion, the term "elementary particle" in the meaning it has been used hitherto, has become obsolete. However, we retain it without being afraid of misunderstandings, even though the adjective "elementary" is not understood in its direct meaning already for a long time.

## Sec. 228. INTERACTION OF FAST ELECTRONS

If electrons move slowly, then the forces of interaction are determined by the arrangement of electric charges at the instant this interaction is being determined. Therefore, in the case of slow electrons, the fact that an electromagnetic field exists is of no significance; it is unessential that the interaction is transmitted through a field.

The situation is quite different in the case of fast particles. Here, it is necessary to take into consideration the lag of interaction due to the finite velocity of propagation of an electromagnetic field. Interaction at the instant  $t$  is determined by

the arrangement of the charges at the instant  $t - \frac{r}{c}$ . Now we cannot manage without considering the field. How can the quantum nature of a field be taken into account in adhering to the interaction scheme: particle—field—particle? This problem is studied in quantum electrodynamics, a branch of theoretical physics still being developed. In considering the interaction of fast particles, experimental evidence compels us to assume that such interaction consists in the transfer of an energy quantum to the field by a particle followed by the subsequent transfer of this quantum to another particle.

If the magnitude of the transferred quantum of energy is equal to  $\mathcal{E}$  and the time required to transfer this quantum from particle to particle is equal to  $\tau$ , then, according to the principle of uncertainty (see Sec. 185), we have

$$\mathcal{E}\tau \sim h.$$

If the particles are close to each other, then  $\tau$  is small and the particles can exchange quanta  $h\nu$  of low and high frequencies. As the distance between particles increases,  $\tau$  increases and an energy exchange becomes possible only by means of small quanta. Reasoning in such a way, one can derive a theoretical formula for the force of interaction between particles.

Many authors describe the quantum interaction of particles through a field in the following picturesque terms: to transfer a quantum of energy  $\mathcal{E}$  to a field, a particle “lends” this energy for a short period of time  $T$ . The uncertainty principle governs the dependence of the lending time on the amount of loaned energy. It is as if the law of conservation of energy were “violated” for the period of time  $T$ , i.e., for the lending time. The uncertainty principle indicates the permissible time interval during which the law of conservation of energy may be “violated” so that this “violation” has no physical meaning.

This viewpoint may be developed further, but it leads to the following difficulty. Since there is no limitation in the absorption and radiation of photons, infinitely large changes in the value of a particle’s intrinsic energy occur when photons are exchanged.

It is interesting that this occurs in all approaches to the evaluation of the intrinsic energy of an electron or some other charged particle. As we know (Sec. 91), a point particle possesses infinite energy. At the same time, it is impermissible to assume that the particle has finite dimensions. In fact, a perfectly solid particle having finite dimensions cannot exist according to the theory of relativity. (The existence of perfectly solid bodies is inconsistent with this theory, since interaction would be propagated instantaneously through such bodies). But if the electron is a deformable particle, what is its structure? A solution to this problem has not yet been found.

A distinctive feature of the new theory is that, in spite of the fact that it is based on contradictions and lacks logical harmony, it leads to a number of new, very interesting results.

## Sec. 229. MESON THEORY OF NUCLEON INTERACTION

In the preceding section we stated that electrically charged particles interact through the medium of an electromagnetic field by means of quanta. An electrically charged particle transmits a quantum of energy to an electromagnetic field and this energy is then transmitted to another particle. If it is assumed that a field is associated with nuclear forces and that this field is also of a quantum nature, then interaction between nucleons can be described as follows: each nucleon is

surrounded by a field; a nucleon transmits a quantum of energy to the field, and the latter transfers this energy to another nucleon.

A theoretical study to determine whether such an explanation of nuclear forces is permissible was undertaken by the Japanese physicist H. Yukawa in 1935. It turned out that a theory could be developed if it is assumed that the field through which nucleons interact possesses quanta having a rest mass that is not equal to zero. Thus, interaction between nucleons was reduced to an exchange of particles having a mass  $m_0 \neq 0$ . For reasons to be explained below, such particles became known as *mesons*. A meson is a quantum of the mesonic field which surrounds nucleons.

Let us dwell on several conclusions to be drawn from the theory. First of all, we shall try to estimate the range of nuclear forces.

The energy of a meson which is transferred by a nucleon during interaction cannot be less than  $m_0 c^2$ , where  $m_0$  is the rest mass of the meson. The order of magnitude of the time taken to transfer a meson is not greater than  $\frac{h}{m_0 c^2}$  (on the basis of the relation  $\mathcal{E}\tau \sim h$ , where  $\mathcal{E}$  is the quantity of transferred energy and  $\tau$  is the time taken to transfer this quantum). Since the velocity of a meson does not exceed  $c$ , a meson cannot be transferred over a distance greater than  $\frac{h}{m_0 c}$ . This constant quantity should characterise the range of nuclear forces.

Such is the conclusion drawn by Yukawa. Using the then known range of nuclear forces, Yukawa showed that theoretical and experimental results coincide when  $m_0$ , the rest mass of a meson, is 200 to 300 times greater than the mass of an electron.

Positive, negative, and neutral mesons have equal validity in this theory. Thus, all interactions between nucleons can be easily interpreted within the framework of this theory. Designating Yukawa's meson by  $\pi$  and using a superscript to indicate the sign of the meson, we can express the interaction between two neutrons or two protons as an exchange process involving a neutral meson:

$$p \rightleftharpoons p + \pi^0, \quad n \rightleftharpoons n + \pi^0.$$

An exchange interaction between a proton and a neutron is a process involving a positive or a negative meson:

$$p \rightleftharpoons n + \pi^+, \quad n \rightleftharpoons p + \pi^-.$$

Yukawa's theory was developed before the discovery of mesons. At present, the above interaction formulas no longer represent theoretical predictions, but rather expressions for phenomena which have been actually observed.

#### Sec. 230. MESONS

The term "meson" (from the Greek "mesos", meaning, 'average', 'intermediate') was coined to indicate that the mass of such a particle lies between that of an electron and that of a proton. Experiments show that several kinds of mesons exist. Mesons (electrically charged) were first detected in cosmic rays. At the present time mesons are produced in accelerators—they arise when nucleons collide.

However, not all mesons play the same role in interactions between nucleons. The Yukawa meson is a  $\pi$ -meson or pion. As has been indicated, there exist positive, negative, and neutral  $\pi$ -mesons. Recent measurements indicate that the mass of a  $\pi$ -meson is equal to  $273m_e$ , where  $m_e$  is the mass of an electron.

The mesons which were first found in cosmic rays are  $\mu$ -mesons or muons (positive and negative). The mass of a muon is equal to  $207m_e$ . Ten years elapsed after the discovery of muons before it was shown that such mesons are products of pion decay. The reason why researchers were able to detect rather easily muons, but were unable to detect pions lies in the different lifetime of these mesons. The average lifetime of a muon is about  $10^{-6}$  second, while that of a pion is about a hundredth of this value, i.e., of the order of  $10^{-8}$  s.

This transformation can be detected by means of photographs. Many of them indicate that the velocity of a pion decreases (this fact is easily recognised by the change in the thickness of the particle track: the slower the particle, the heavier the track—since a slow particle creates more ions along its path than a fast particle). At the point where the track ends, another track begins. Calculations show that the new particle is of very high energy, which may be explained only if we assume that a part of the rest energy of a pion has been transformed into kinetic energy. Furthermore, in order not to violate the law of conservation of momentum, we must assume that a neutral particle of very small mass is created when a pion is transformed into a muon. Here, we encounter the neutrino once again, but this neutrino is of another kind. Successful results of fine investigations have shown that pions are transformed into photons or a pair of leptons according to the following schemes:

$$\pi^+ \rightarrow \mu^+ + \nu_\mu, \quad \pi^+ \rightarrow \bar{\nu}_e + e^+, \quad \pi^- \rightarrow e^- + \bar{\nu}_e, \quad \pi^- \rightarrow \mu^- + \bar{\nu}_\mu.$$

Thus, there exist two kinds of neutrinos and two kinds of antineutrinos.

The rest mass of a pion is about 150 MeV. Therefore, a pion can be produced by nuclear bombardment with projectiles having an energy greater than this value. Actually, an energy higher than 300 MeV is required. Nuclear bombardment with high-energy particles results in the creation of a large number of mesons.

It is particularly difficult to prove the occurrence of neutral pions, since the lifetime of such a particle turned out to be equal to  $10^{-15}$  second. This means that the particle traverses a distance of only a thousandth of a millimetre during its lifetime.

We are not going to dwell on other mesons detected in an ever increasing number as the capacity of particle accelerators increase. Pions may be regarded as the principal state of the entire family of mesons which will be given our consideration after a study of the baryon spectrum.

#### Sec. 231. RELATIVISTIC THEORY OF AN ELECTRON

In discussing the Schrödinger equation (see Sec. 184), we did not resort to the relations of the theory of relativity, since we assumed that the velocities of the particles were much less than the velocity of light.

But we cannot dispense with the relativistic corrections when dealing with high-energy electrons, i.e., electrons having energies of the order of millions of electron volts. Such high-energy electrons are encountered in radioactive radiation, in X-ray tubes whose operating voltages are equal to millions of volts, in betatrons, etc.

To correctly describe electrons having such velocities by a wave function, we must take into account the relationship between energy and momentum given by the theory of relativity.

If the rest mass of a body is given, the energy and momentum of the body can be directly determined from its velocity. Therefore, energy and momentum are related.

Squaring the expressions

$$\frac{\mathcal{E}}{c} = \frac{m_0 c}{\sqrt{1-\beta^2}}, \quad p = \frac{m_0 v}{\sqrt{1-\beta^2}}$$

and then subtracting, we get

$$\frac{\mathcal{E}^2}{c^2} = p^2 + m_0^2 c^2.$$

This relationship can be written in the following form:

$$\mathcal{E} = \pm c \sqrt{p^2 + m_0^2 c^2}$$

(see Fig. 253). For  $p = 0$   $\mathcal{E} = \pm m_0 c^2$ . This means that the coordinates of the points at which the two branches of the given curve intersect the axis represent equal rest energies of the particle.

For a long time no attention was paid to the negative branch of the curve. It seemed impossible that particles could exist which obeyed the laws of this branch.

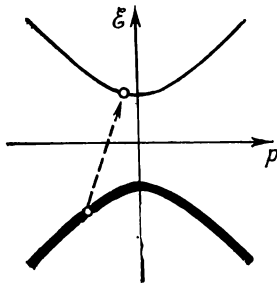


Fig. 253

Indeed, a particle having negative rest energy and whose energy decreases as the momentum increases must behave very strangely. It is like a particle having negative mass, which means that when a force is exerted on such an "electron" the particle is accelerated in the opposite direction to that of the force. Thus, if we try to attract such a particle, it is repelled. Imagine that we have two close electrons—an ordinary one and an extraordinary one, i.e., one which obeys the laws of the lower curve. Electrons should repel each other. Therefore, the ordinary electron will tend to move away from the extraordinary one. But since the latter "electron" will be attracted under the action of the force of repulsion, it will move to-

ward the ordinary electron. As a result, this pair will be accelerated jointly, whereby the increase in positive kinetic energy of the ordinary electron will be counterbalanced by the increase in negative energy of the extraordinary particle.

Does such a particle exist? First of all it should be noted that from the viewpoint of quantum mechanics this is not another particle at all. The negative branch of the energy-momentum curve should be interpreted as giving the lowest energy level of one and the same particle, namely, an electron. An electron may be located at an ordinary energy level or at a negative energy level. Like in the other cases, transition to the lowest energy level should be accompanied by energy radiation.

To be sure, in such a transition the radiation cannot be limited to a single photon. This can be explained as follows. When an electron passes from an  $\mathcal{E}_1, p_1$  state of the upper curve to an  $\mathcal{E}_2, p_2$  state of the lower curve, the energy released is equal to

$$\mathcal{E}_1 - \mathcal{E}_2 = c (\sqrt{p_1^2 + m_0^2 c^2} + \sqrt{p_2^2 + m_0^2 c^2}),$$

that is more than  $c(p_1 + p_2)$ . If this energy were transferred to a single photon, then we should have  $\mathcal{E}_1 - \mathcal{E}_2 = h\nu$ , and, therefore,  $p_1 + p_2 = \frac{h\nu}{c}$ , the change in the momentum of the photon ( $p_1$  is anti-parallel to  $p_2$ ) would equal  $\frac{h\nu}{c}$ . But  $\mathcal{E}_1 - \mathcal{E}_2$  is greater than  $c(p_1 + p_2)$ . In the case of a single photon it is not pos-



sible to satisfy simultaneously the law of conservation of energy and the law of conservation of momentum, but this is always possible if two photons are emitted.

We should not be disturbed by the fact that the lower state corresponds to negative energy and that particles with  $\mathcal{E} < 0$  behave very strangely. The difficulty lies elsewhere: since there is limitless room at the lower level in view of the fact that it extends to all possible values of momentum, why do all ordinary electrons not pass over to this level? The proposed model accounts for the existence of only "extraordinary" electrons. To accord with this model, an ordinary electron should have a short lifetime, much like an excited atom.

## Sec. 232. CREATION AND ANNIHILATION OF PAIRS OF PARTICLES

A relativistic theory of the electron was developed by P.A.M. Dirac in 1927. The description of an electron in this theory differs considerably from that in the Schrödinger theory. Four wave functions are used to describe the behaviour of an electron, since one wave function no longer suffices. The existence of electron spin does not follow from the Schrödinger theory, but electron spin is a necessary consequence of the Dirac theory. The success of this theory in describing numerous phenomena speaks for the validity of its basic concepts. Here we shall consider in the light of this theory only the contradiction discussed in the preceding section. How can we explain the fact that a tremendous number of ordinary electrons exist? How can we explain the "reluctance" of these electrons to pass over to the lower, negative energy level?

The Pauli exclusion principle does not allow an electron to pass over to a lower level if that level is occupied by two electrons having opposite spins. Let us apply this principle to our problem. The solution will be obtained if we assume that all negative energy states are occupied. Does this mean that all space is completely filled with electrons having a negative energy state? This conclusion seems to be inevitable. The Dirac theory has led to a new view of vacuum. In this theory, vacuum acquires physical properties, that is, it is filled with electrons in a negative energy state. Moreover, it is filled boundlessly, since the decrease in energy can extend to negative infinity.

Let us see which phenomena can be explained and predicted by this theory. If all negative states are filled, their existence may be detected only when an electron passes from a negative energy level to a positive energy level after receiving a significant portion of energy—in any case, not less than  $2m_0c^2$ . Such a process may involve the expenditure of two photons. Another possibility is the following: a photon passing close to a heavy atomic nucleus, which has a strong electric field, may give up its energy to raise an electron from a negative to a positive level. The role of the atomic nucleus consists in providing the necessary momentum.

In both cases an electron is "given birth". But in addition to the creation of an electron, a "hole" appears in the negative energy states. The removal of an electron, i.e., a negative charge, means that the hole acquires a positive charge of equal magnitude. On the other hand, the absence of a particle having negative energy signifies that the energy has increased. Hence, the greater the momentum of a hole, the greater its energy. Here we come to a conclusion that "holes" behave like positively charged electrons having positive energy. Except for the sign of the charge, the behaviour and laws of motion of a positive electron (positron) in no way differ from those of an ordinary electron.

A positron and an electron are "born" as twins at the expense of the energy of photons. The reverse process, annihilation, is also possible. This consists in the

transformation of a colliding electron and positron into two photons or, if annihilation occurs close to a heavy atomic nucleus, into a single photon.

It is clear why positrons have a short lifetime: they are attracted by electrons and disappear upon collision. But why then do electrons not disappear? The reason is simply that there is an excess of them.

Do systems exist in which there are an excess of positrons, that is, in which electrons are unstable? Such a supposition is not at all ridiculous.

Pair creation and annihilation are processes which can be easily observed in large numbers under laboratory conditions. When gamma rays whose energy exceeds 1 MeV pass through a thin metal foil, the electrons which are ejected from the foil will be deflected in a magnetic field in opposite directions. By tracing the path of a positron (means of observing charged particles are discussed in Sec. 211), we can determine the point at which the particle disappears. This is where the positron combined with an electron. With the aid of modern photon counters, we can show that two oppositely directed photons, each having an energy of the order of  $\frac{Mc^2}{2}$  eV, simultaneously emerge from this spot.

### Sec. 233. PARTICLES AND ANTIPARTICLES

There is no reason for supposing that the existence of the positron (it would be better to call it now an antielectron) is due to a peculiar feature of small particles. In spite of the distinctive features of the theory of interaction between nucleons, it is basically similar to the theory of interaction between electrons. In most theoretical studies, nucleons are suggested to be described by equations which are quite similar to the Dirac equations for electrons. It was to be expected, therefore, that nucleons have antiparticles which stand in the same relationship to the proton and neutron as a positron does to an electron. The first of such antinucleons to be discovered was the antiproton. Somewhat later the antineutron, whose magnetic moment orientation differs from that of the neutron, was discovered. (The magnetic moment and the angular momentum vector are antiparallel in the case of the neutron and parallel in the case of the antineutron.)

The discovery of the antiproton proved the correctness of the general idea of an inseparable link between field and particles. Like in the case of a positron-electron pair, a proton-antiproton pair can be born by the passage of a nucleon from a negative energy state to a positive energy state. For this purpose, an energy of at least  $2Mc^2$  is required. This is a tremendous amount of energy—1,840 times the energy required to create an electron-positron pair. An accelerator which could accelerate particles to billions of electron volts had to be constructed before it was possible to discover the antiproton.

A collision of a proton with an antiproton results in their annihilation. Since nucleons transfer energy through a meson field, as the annihilation proceeds, their mass and energy are given up to quanta of this field, that is, to mesons. This process is being carefully studied.

Figure 254 shows a photograph of the annihilation of a proton and an antiproton. This process was observed in a bubble chamber filled with liquid propane. A scheme of the process is given in the upper left-hand corner.

The reasoning used in explaining why antiparticles must exist extends to the neutrino as well. Its "mirror" image is called the antineutrino. The difference between the particles composing this doublet is the same as in the case of the neutron-antineutron doublet.

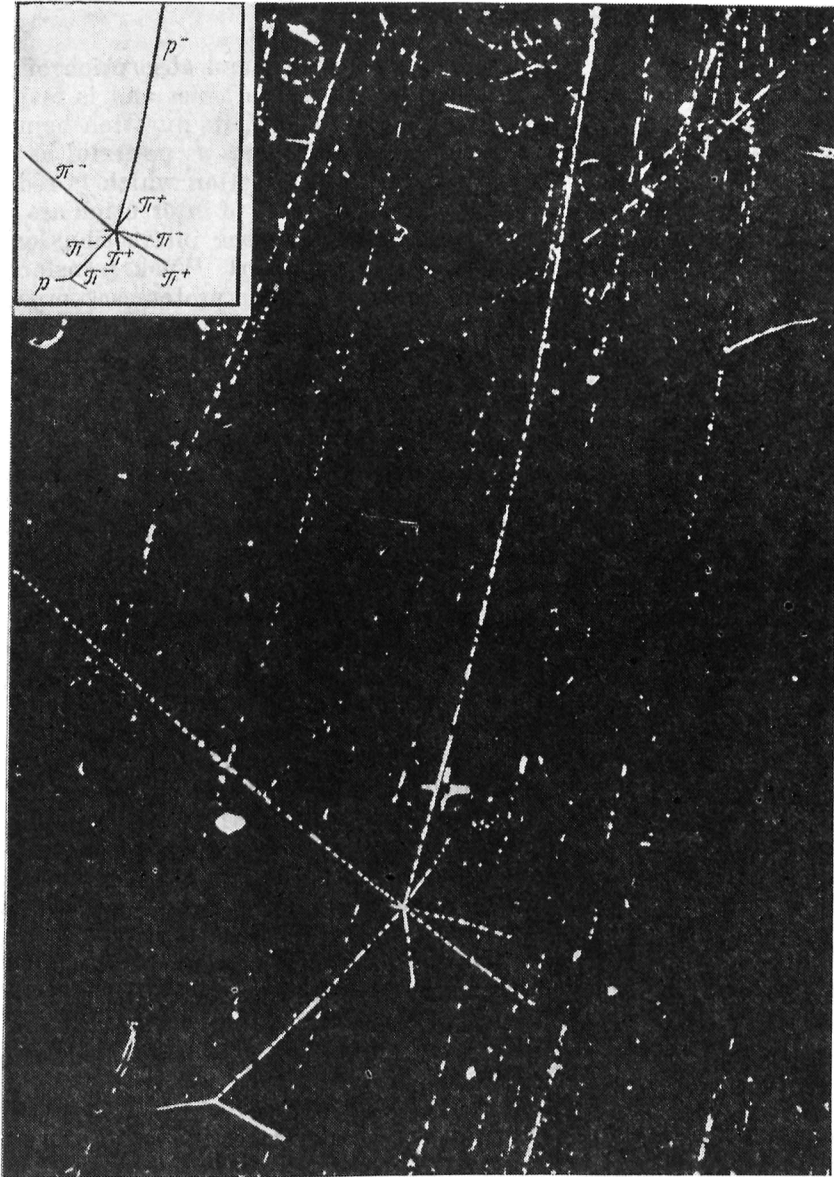


Fig. 254

Muons and other elementary particles which have not been discussed are also encountered as doublets.

Muons constitute a triplet: there is a muon having a positive charge, another having a negative charge, and still another having a zero charge. In contrast to the neutron and neutrino, the neutral muon has no spin and, therefore, can have no antiparticle (in other words: the neutral muon coincides with its antiparticle). The photon is another particle which has no "mirror image".

## Sec. 234. ASYMMETRY OF ELEMENTARY PARTICLES

Interactions of nucleons is accompanied by emission and absorption of muons. This is the strongest interaction between elementary particles and is responsible for the force binding the nucleons in an atomic nucleus, its duration being equal to  $10^{-23}$  second. Nuclear forces are about a hundred times as powerful as electromagnetic forces; the duration of electromagnetic interaction which is reduced to the exchange of photons is  $10^{-21}$  second. These two types of interaction are conventionally called *strong* interactions in contrast to the *weak* interactions occurring in particle transformations in which neutrinos take part. Weak transformations are exemplified by the transformation of a neutron into a proton, accompanied by the emission and absorption of a neutrino and an electron (see the above discussed  $\beta$ -decay), and by the decay of  $\pi$ - and  $\mu$ -mesons:

$$\pi \rightarrow \mu + \nu, \quad \mu \rightarrow e + \nu + \bar{\nu}.$$

The duration of weak interactions is of the order of  $10^{-9}$  second. Nuclear forces in such processes are about  $10^{-14}$  times as powerful as weak interaction forces. It goes without saying that in estimating the interaction force by the duration of the process we are assuming that other conditions remain equal.

Some extremely interesting discoveries have been made recently in connection with weak interactions. It has been found that weak processes exhibit "left-right" asymmetry. Thus, for example, in beta-decay of cobalt nuclei polarised at low temperatures by means of a magnetic field (polarisation of the particles consists in orientation of their magnetic moments and spins in a definite direction), the angular distribution of electrons is asymmetrical with respect to the "forward" and "backward" directions. Similarly, in  $\mu$ -meson decay, asymmetry with respect to the direction of motion of the particles has been detected.

A theory for this phenomenon was proposed by the American scientists C. N. Yang and T. D. Lee, and also by the Soviet physicist L. D. Landau. The phenomena under consideration may be explained in two different ways: either by internal asymmetry of the particles or by asymmetry of space. We shall confine ourselves to the first explanation. Its essence lies in the assumption that elementary particles are like screws as regards their symmetry properties. Such asymmetrical particles are well known to the physicist. They include, for example, the right and left optical antipodes of molecules (discussed in detail in Sec. 154). It is clear that an asymmetrical elementary particle whose axis is randomly oriented has different properties in the "forward" and "backward" directions. This has been proved experimentally.

In order to explain the asymmetry observed in experiments with muon beams, use can be made of the Landau hypothesis, which relates the asymmetry of a particle to its charge. As indicated in the preceding section, all particles (except for the photon) are encountered in nature as charged pairs. Landau proposed that if the symmetry properties of a particle are like those of a right-handed screw, the symmetry properties of an antiparticle are like those of a left-handed screw. Reflection in a mirror makes the right hand appear like the left hand and a right-handed screw like a left-handed screw. According to the suggested hypothesis, the "mirror image" of a particle is its antiparticle.

What application has this hypothesis to experiments with muon beams? It can be proved that a particle having no mass must be oriented with its spin in the direction of motion. The mass of a neutrino is, obviously, equal to zero. Therefore, all neutrinos are "longitudinally polarised". The difference between a neutrino and an antineutrino is reduced to the following: the spin of a neutrino is oriented

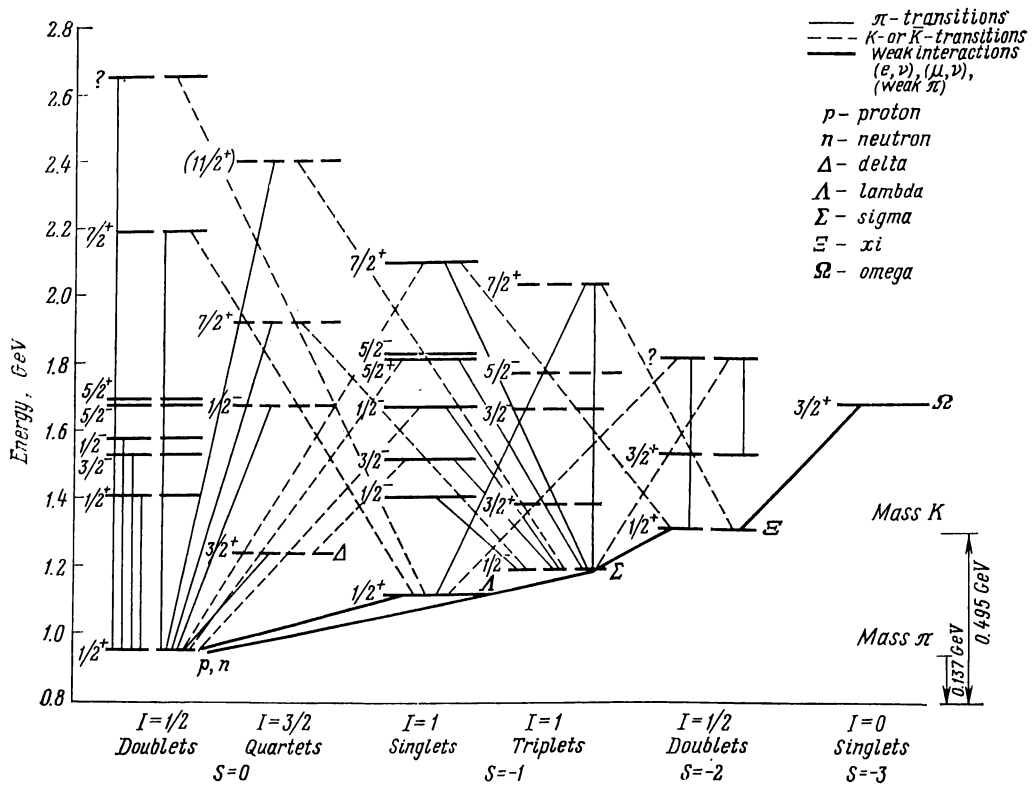


Fig. 255

in the direction of its motion, while that of an antineutrino is in the opposite direction. Muons are formed when pions decay. But the spin of a pion is equal to zero, therefore the spin of a muon must be parallel to the spin of a neutrino, i.e., the muons in such a beam will be polarised longitudinally. This explains the asymmetry observed in the distribution of electrons during the subsequent decay of these muons.

The particles possessing in their motion the symmetry properties of a right- and left-handed screws are said to be particles of opposite parity, and are designated by the signs (+) and (-), respectively.

### Sec. 235. BARYON SPECTRUM

When colliding with other particles, nucleons become excited and pass into a large number of different quantum states. A nucleon in an excited state is called a baryon.

Our today's knowledge of the system of baryon levels and transitions between them is represented by the diagram suggested by the American physicist V. Weisskopf (see Fig. 255). The heavy horizontal lines indicate the energy levels of baryons found experimentally, the energy scale is given on the left. For the energy levels of the atom nuclei discussed in the preceding section we need a scale hundreds of thousand times larger than that used for atoms.

Passing to the spectrum of baryons, we have to magnify the scale another thousand-fold. We see that the energy differences between the levels are measured here in larger units, viz., in GeV (gigaelectronvolts). This scale alone shows the purpose of ever increasing powers obtained in particle accelerators: the realisation of this diagram becomes possible due to the latest experimental work on accelerators which impart to projectile particles energies sufficient for a baryon to get excited.

The whole diagram should be understood as a spectrum pattern of one particle the ground state of which (nucleon) is a doublet. The energy difference between the components of this doublet, i.e., between a proton and neutron, equal to 1.2 MeV is not seen on the given scale.

For the sake of obviousness, baryon levels are grouped in columns differing in two quantum numbers—the isospin  $I$  and the strangeness  $S$ .

As it has been noted, a difference in the charges displaces the level by a magnitude imperceptible in the scale to which our diagram is drawn. Experiments show that some states are encountered in a single charge species called singlets. The third column from the left (as also the last one) represent singlet levels. The lowest singlet level of a baryon is known as a lambda particle; this particle is electrically neutral.

Located above the principal proton-neutron level (the first column) are other doublets. The second column represents quartets. The lowest level of this family—the delta particle ( $\Delta$ )—is encountered in four charge variants:  $\Delta^-$ ,  $\Delta^0$ ,  $\Delta^+$  and  $\Delta^{++}$ . The fourth column contains triplets, the fifth doublets, and the sixth singlets.

The isospin number  $I$  is assigned to the states of one and the same multiplicity (the term “isospin number” is extremely infelicitous since this number has nothing to do with rotary spin). It is chosen so that  $2I + 1$  is equal to the multiplicity.

The levels are also grouped into columns depending on the strangeness  $S$ . The strangeness  $S = Y - A$ , where  $A$  is the baryonic number. For the table under discussion  $A = +1$ . In the case of antibaryons,  $A = -1$ . Mesons have a baryonic number equal to zero.

The baryonic number of atomic nuclei is equal to the number of the nuclei. The quantity  $Y$  is equal to the doubled average charge of a multiplet:

$$\begin{array}{ll}
 \text{the first column } Y = 2 \cdot \frac{1}{2} (0 + 1), & S = 0; \\
 \text{the second column } Y = 2 \cdot \frac{1}{4} (-1 + 1 + 0 + 2), & S = 0; \\
 \text{the third column } Y = 2 \cdot \frac{1}{2} \cdot 0, & S = -1; \\
 \text{the fourth column } Y = 2 \cdot \frac{1}{3} (-1 + 0 + 1), & S = -1; \\
 \text{the fifth column } Y = 2 \cdot \frac{1}{2} (-1 + 0), & S = -2; \\
 \text{the sixth column } Y = 2 \cdot (-1), & S = -3.
 \end{array}$$

The energy levels of one and the same column (with equal  $S$  and  $I$ ) differ in the values of spins and parity.

Let us now consider the transitions between the levels. The scheme indicates  $\pi$ -transitions (continuous lines),  $K$ -transitions (dash lines), and transitions accompanied by emission of a lepton pair (heavy continuous lines). Photon emission is not shown in the figure. Photons are usually emitted in pion transitions, provided

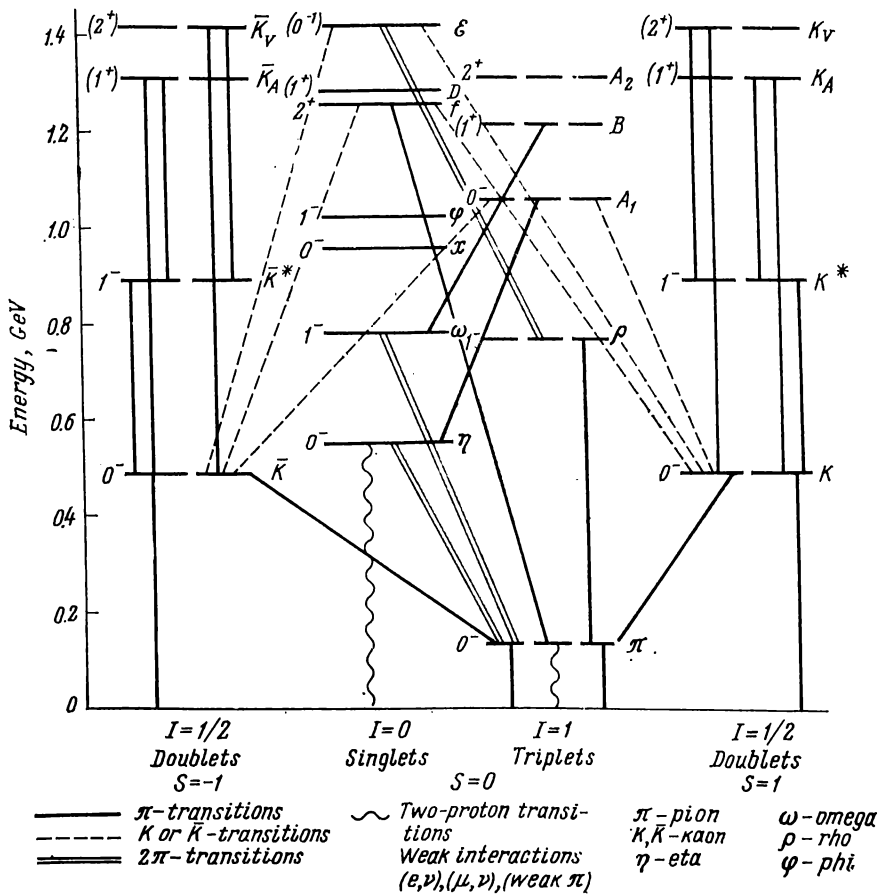


Fig. 256

the charge remains unchanged. Transitions are possible from each member of one multiplet to each member of the other. In order not to overload the drawing, we have confined ourselves to one line.

Pion transitions are possible only between the levels with equal values of  $S$ . Just this strangeness in baryon behaviour was the reason for introducing the number  $S$ .

If we assign a zero strangeness to all pions, and in the case of kaons, set  $S = +1$  for  $K^+$  and  $K^0$  and  $S = -1$  for  $\bar{K}^0$  and  $\bar{K}^-$  (a letter with a dash denotes an antiparticle), then the law of conservation of strangeness will take place.

If the isospin of a pion is equal to 1 and that of a kaon to  $\frac{1}{2}$ , then in transitions the isospin number is retained as well.

Despite the fact that mesons are considered in this scheme as peculiar emission quanta, it turns out to be expedient to treat the whole family of mesons as an excited state of a pion. However, one should not forget that there is an essential difference between a pion and a nucleon. Pions are unstable and transform into photons or lepton pairs:  $\pi^0$  is transformed into photons during  $10^{-16}$  second, charged  $\pi$ -mesons into lepton pairs within  $10^{-8}$  second.

Figure 256 represents a scheme for a meson spectrum. As can be seen from the diagram, it is arranged according to the same principle, resembling the picture for the baryon spectrum. But this scheme is not so "beautiful" as the preceding one. We see that kaons and eta-mesons are capable of disappearing without passing through a pion state. Obviously, muons are not seen in this scheme, since they are members of lepton pairs.

### Sec. 236. QUARKS

The spectra of the atom and atomic nucleus are explained by dynamics and interaction of their components—electrons and a nucleus in the case of an atom, and a proton and neutron in the case of an atomic nucleus.

We think that the baryon and meson spectra can be explained in the same way. If we succeeded in showing that nucleons and mesons are made up of certain more elementary particles and in explaining the level systems of baryons and mesons by interaction of these subelementary particles, then elementary-particle physics would acquire the completeness of quantum mechanics. But at present, this problem is far from being solved.

However, recently a simple hypothesis explaining the structure of baryons and mesons was suggested. It describes the observed energy levels so successfully that it cannot be regarded as a random success. The hypothesis has no experimental evidence yet. Either it will be obtained with the aid of newly designed, more powerful accelerators, or the hypothesis will prove groundless.

Even in the absence of experimental evidence the quark scheme will remain useful as a method of classification of the energy levels of baryons and mesons. To explain the spectra in question three types of quarks are introduced:  $p$  (charge  $+\frac{2}{3}$ , strangeness 0),  $n$  (charge  $-\frac{1}{3}$ , strangeness 0), and  $\lambda$  (charge  $-\frac{1}{3}$ , strangeness  $-1$ ). The corresponding antiparticles are also needed; they are obtained by reversing the signs of strangeness and charge. Each quark has a spin equal to  $\frac{1}{2}$ .

The baryon and meson spectra shown in Figs. 255 and 256 are explained in an unbiased way. To this end, we have to assume that a baryon consists of three quarks, and a meson of two (quark + antiquark). It is easy to check that correct values of strangeness, charge, and spin are obtained immediately if

$$\begin{array}{ll}
 \text{neutron} = 2n + p & \uparrow\uparrow\downarrow \\
 \text{proton} = 2p + n & \uparrow\uparrow\downarrow \\
 \Delta^- = 3n & \uparrow\uparrow\uparrow \\
 \Delta^0 = p + 2n & \uparrow\uparrow\uparrow \\
 \Delta^+ = 2p + n & \uparrow\uparrow\uparrow \\
 \Delta^{++} = 3p & \uparrow\uparrow\uparrow \\
 \Lambda = p + \lambda + n & \uparrow\uparrow\downarrow
 \end{array}$$

Pions have to be constructed from the pairs  $p + \bar{n} \uparrow\downarrow$  ( $\pi^+$ ),  $p + p^-$  or  $n + \bar{n} \uparrow\downarrow$  ( $\pi^0$ ) and  $n + p$  ( $\pi^-$ ), and so on. This scheme enables us to describe not only the structure of baryons and mesons, but also the transitions between the levels.

Of course, the introduction of quarks, even if it proves a success, will solve not all the problems of elementary-particle physics. On the whole, this field of knowledge is not yet so orderly and clear as other branches of physics.



## Atomic Structure of Bodies

## Sec. 237. POLYCRYSTALLINE SUBSTANCES AND MONOCRYSTALS

As a rule, crystals begin to grow around a very large number of centres in a melt or solution. If special measures are not adopted, a polycrystalline substance rather than a monocrystal will form as the result of crystallisation. Under a microscope such a substance seems to consist of individual grains (see Fig. 257). Each grain is a crystal which has an irregular haphazard form due to the fact that its normal growth has been impeded by neighbouring crystals. Most bodies commonly encountered, particularly metals and rocks, are polycrystalline substances.

The boundary between grains is revealed by etching with an appropriate solvent. This is due to the fact that most of the impurities of a substance accumulate at the grain boundaries. The interlayer between crystals differs from the "body" of the grain not only in that it contains foreign atoms, but in that its atoms have

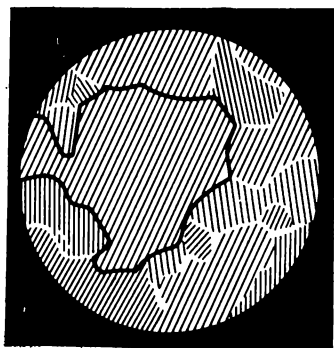


Fig. 257

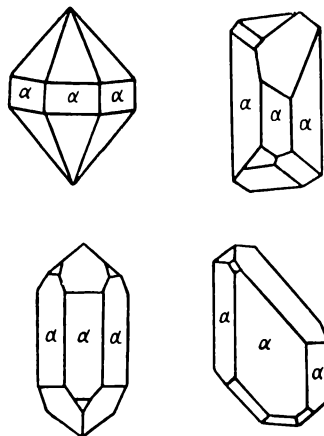


Fig. 258

a disturbed (transitional) arrangement. The basic structure of the boundary between grains is clearly visible under a microscope as peculiar, smooth "paths". The usual size of grains in metals and rocks is  $10^{-4}$ - $10^{-5}$  cm.

A single crystal (monocrystal) of any crystalline substance may be found in nature or artificially grown. A monocrystal is distinguished by its regular shape, i.e., plane faces, straight edges, and symmetry, in other words, the proportionality of its component parts. This regular shape reflects a crystal's internal properties, which enable us experimentally to distinguish a crystal from a bit of material given such a shape artificially. It is also not difficult to recognise a crystal when its characteristic features are hidden. Thus, a sphere may be fashioned from a large crystal of rock salt but, when it is placed in water, surface material is dissolved at a non-uniform rate and, as this process goes on, its accidental shape tends to be transformed into the polyhedral shape which is natural for this substance. A monocrystal can be easily distinguished from a polycrystal by means of X-ray analysis.

A naturally formed crystal has the shape of a polyhedron. As in the case of every

polyhedron, a crystal has a certain number of faces ( $p$ ), edges ( $r$ ) and corners ( $e$ ), which are related to one another as follows:  $p + e = r + 2$ . For example, a cube has 6 faces, 8 corners and 12 edges.

Crystal faces are arranged in bands or zones. A system of faces the intersections of which are parallel edges is called a *zone*, and the direction of these edges is called *the axis of the zone*.

Crystals of one and the same substance may differ considerably with respect to shape, but it has been long known that a given substance has characteristic angles between faces and edges. (Depending on chance, one component of a crystal may grow more than another; as a result, apparently, the proportionality between components may be upset.) This important rule, which is sometimes called *the law of constant angles*, is illustrated by Fig. 258. In the figure, we see four different crystals of silicon dioxide ( $\text{SiO}_2$ ). It is seen that the number of faces and their relative dimensions differ from specimen to specimen, but the angles between corresponding (i.e., related by one and the same element of symmetry and denoted in the figure by the Greek letter  $\alpha$ ) faces and edges remain unchanged.

### Sec. 238. SPACE LATTICE

The distribution of matter in a crystal may be represented by a three-dimensional periodic function. This is the fundamental rule lying at the basis of crystal investigations.

Fig. 259 shows a wall-paper pattern. A certain element of this pattern is repeated in two directions. Consider any point  $A$  in the figure. A system of lines may be

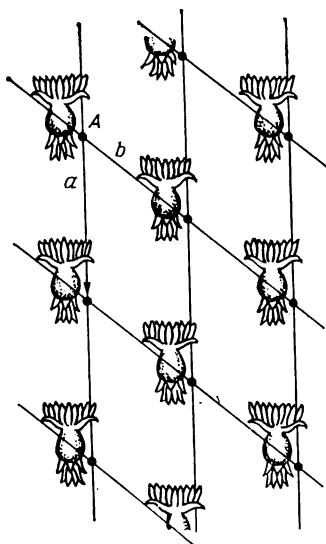


Fig. 259

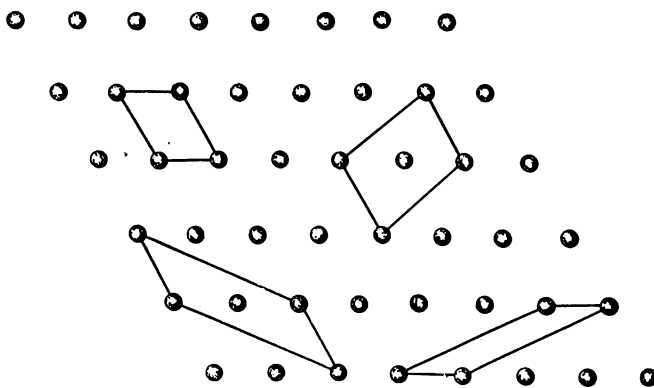


Fig. 260

drawn through the selected points (*nodes*) as shown. A pattern element the repetition of which yields the full pattern is enclosed within a cell of the resulting lattice. Evidently, the entire pattern can be obtained from a single cell by means of parallel translations of the cell vectors  $a$  and  $b$ .

A crystal constitutes a space lattice, not a plane lattice. An element of a crystal is a parallelepiped based on three translational vectors  $a$ ,  $b$ ,  $c$ , which may be

selected, generally speaking, in an infinite number of ways. Such a parallelepiped will be called an *elementary* or *unit cell*, the vectors  $\mathbf{a}$ ,  $\mathbf{b}$ ,  $\mathbf{c}$  the *basic translational vectors*, or simply *basic vectors*, and their lengths  $a$ ,  $b$ ,  $c$  the *basic repetition periods* or *lattice spacings*. The lattice is described in a system of coordinates the axes of which coincide with the directions of the basic vectors. Different ways of selecting basic vectors, i.e., an elementary cell, are illustrated for a two-dimensional case in Fig. 260. An elementary cell in the general case is an oblique-angled parallel-

epiped with edges  $a$ ,  $b$ ,  $c$  and angles  $\alpha = \angle b, c$ ,  $\beta = \angle c, a$ ,  $\gamma = \angle a, b$ . The six quantities which uniquely describe an elementary cell are called its *parameters*. Since the entire lattice is determined when an elementary cell is given, the above quantities are sometimes called *the parameters of the lattice*.

A cell in the form of an oblique-angled parallelepiped is said to be *triclinic* and if  $\alpha = \gamma = 90^\circ$ , *monoclinic*. A cell in the form of a right-angled parallelepiped is said to be *rhombic* and if in addition  $a = b$ , *tetrahedral*. If  $a = b \neq c$ ,  $\alpha = \beta = 90^\circ$ , and  $\gamma = 120^\circ$ , a cell is said to be *hexagonal*. The simplest cells have the form of a cube.

If one of the lattice points is selected as the origin of the coordinate system, the radius vector of any other lattice point is given by the formula

$$\mathbf{R}_{mnp} = m\mathbf{a} + n\mathbf{b} + p\mathbf{c},$$

where  $m$ ,  $n$ ,  $p$  are whole numbers representing the coordinates of these nodes. The indicated numbers are called *the indexes of the nodes*. The set of three indexes describing lattice points is designated by *the nodal symbol*  $[[mnp]]$ .

There are an infinite number of nodal lines and nodal planes. Nodal lines and nodal planes are represented in a lattice by infinite families of parallels. The transition from one line to another of the same family, or from one plane to another, occurs by translation along a vector joining two nodes of these lines or planes. Each family of nodal lines is described by the lattice spacing along a nodal line and the direction, i.e., incline to the selected coordinate axes. To describe a family, we select the line passing through the origin of the coordinate system. A nodal line is described uniquely by the indexes  $u$ ,  $v$ ,  $w$ , of the first lattice point lying on this line. The indexes of this lattice point are called *the indexes of the line* and are designated by  $[uvw]$ . If an index is negative, a minus sign is placed above the numeral. The symbol  $[100]$  represents the  $a$ -axis of the lattice,  $[010]$  the  $b$ -axis, and  $[001]$  the  $c$ -axis. The lines  $[011]$  and  $[0\bar{1}\bar{1}]$  represent plane diagonals in the face  $bc$ . Of course,  $[011]$  and  $[0\bar{1}\bar{1}]$  are one and the same line. Distinguishing between these two designations has significance only if we wish to emphasise the polarity of the direction. The spatial diagonals of a cell have the symbols  $[111]$ ,  $[1\bar{1}\bar{1}]$ ,  $[\bar{1}1\bar{1}]$  and  $[\bar{1}\bar{1}1]$ . There are four of them, corresponding to the existence of eight quadrants; the other four symbols represent the same lines, but with reverse polarity. Thus,  $[\bar{1}\bar{1}1]$  is anti-parallel to  $[1\bar{1}\bar{1}]$ , etc.

A space lattice can be constructed as follows. First, an infinite plane-lattice (nodal plane) is formed by means of two translational vectors; then, a space lattice is formed by means of a third translational vector which does not lie in this plane. A crystal lattice can be represented by families of nodal planes in an infinite number of ways. Every family of nodal planes consists of parallel planes separated from one another by equal distances. For a given lattice, specification of the interplanar distance and the orientation of one of the planes relative to the selected coordinate axes completely describes a family of nodal planes. It is also sufficient to give the orientation relative to the selected axes of the plane closest to the ori-

gin. The distance of this plane from the origin will be equal to the interplanar distance of the given family.

Let this plane intersect the lattice axes at the coordinates  $\frac{a}{h}$ ,  $\frac{b}{k}$  and  $\frac{c}{l}$ , i.e., fractions of the basic lattice spacings. The numbers  $h$ ,  $k$  and  $l$ , which describe the orientation of the plane, will be called *the indexes of the plane*. It is easily seen that  $h$ ,  $k$  and  $l$  are whole numbers. One way of showing this is as follows. Consider a plane passing through an initial lattice point and another plane, of the same family, displaced by an amount  $a$ . This is shown in Fig. 261. Other planes will pass through these nodal planes, but they must be separated from one another by equal distances. Therefore, the repetition periods along the selected axes will be divided by the nodal planes into a number of equal parts.

The plane closest to the origin and intersecting the axes at  $\frac{1}{h}$ ,  $\frac{1}{k}$  and  $\frac{1}{l}$  of the lattice spacings is described by the set of three indexes  $h$ ,  $k$  and  $l$ . Its symbol is

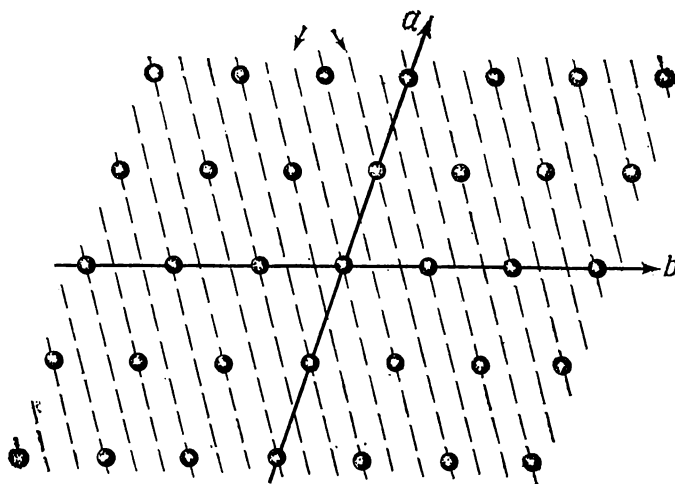


Fig. 261

designated by enclosing these indexes between round brackets:  $(hkl)$ . For example, the plane  $(236)$  intersects the axes at the coordinates  $\frac{a}{2}$ ,  $\frac{b}{3}$  and  $\frac{c}{6}$ . Any plane which intersects the axes at coordinates which are a multiple of these values is a member of this family. Thus, in the case under consideration, the successive planes beyond the one closest to the origin will intersect the axes at the following coordinates:  $a, \frac{2}{3}b, \frac{c}{3}$ ;  $\frac{3}{2}a, b, \frac{c}{2}$ , etc.

If a plane intersects the axes at negative coordinates, this is indicated by a minus sign above the corresponding index. It is evident that the planes  $(hkl)$  and  $(\bar{h}\bar{k}\bar{l})$  belong to the same family. Therefore, all the signs of the indexes of a plane may be reversed.

If a plane is parallel to a coordinate axis, the corresponding index is equal to zero. Thus,  $(110)$  is a plane that is parallel to the  $c$ -axis,  $(001)$  is the lattice plane  $ab$  (see Fig. 262),  $(010)$  is the plane  $ac$ , and  $(100)$  is the plane  $bc$ . Planes passing through one of the axes and one of the diagonals have indexes consisting of two ones and one zero. For example, the plane  $(101)$  is a plane which is parallel to the

$b$ -axis and passes through the diagonal extending from the terminal of vector  $+a$  to that of vector  $+c$  (not the diagonal passing through the origin). The plane  $(\bar{1}0\bar{1})$ , which passes through the terminals of the vectors  $-a$  and  $-c$ , belongs to the same family. The plane  $(10\bar{1})$  and its "reverse side"  $(\bar{1}01)$  are also parallel to the  $b$ -axis and pass through the diagonals  $ac$  which do not begin from the zero lattice point, but extend from the terminal of vector  $+a$  to that of vector  $-c$  and the terminal of vector  $-a$  to that of vector  $+c$ , respectively.

A symbol consisting of three units refers to planes passing through three diagonals. These planes pass through the terminals of all three lattice vectors. Thus, the plane  $(\bar{1}\bar{1}\bar{1})$  passes through the terminals of the vectors  $-a$ ,  $-b$  and  $+c$ .

The indexes of a family of nodal planes are at the same time the indexes of crystal faces. Two parallel faces have the indexes  $(hkl)$  and  $(h\bar{k}\bar{l})$ .

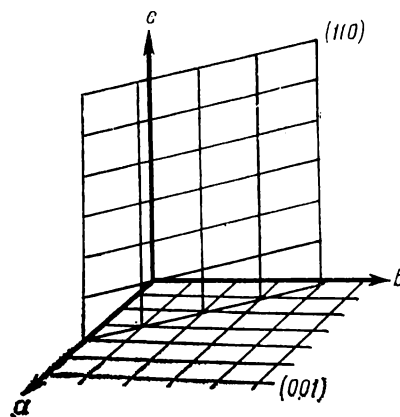


Fig. 262

### Sec. 239. CELL SELECTION AND CRYSTAL SYMMETRY

The vectors  $a$ ,  $b$  and  $c$  of a lattice may be selected in a variety of ways. If there are no lattice points within an elementary cell, we call it a primitive cell.

Various ways of selecting a primitive elementary cell are shown in Fig. 263. In view of the periodicity of a space lattice, the volume associated with each lattice point is constant and is equal to the volume of a primitive elementary cell, regardless of the manner in which such a cell is selected. Since each of the eight lattice points at the corners of such a cell is "shared" by eight cells,  $\frac{1}{8}$  of each of the lattice points belongs to the given cell. Thus, on the average, there is one lattice point per cell.

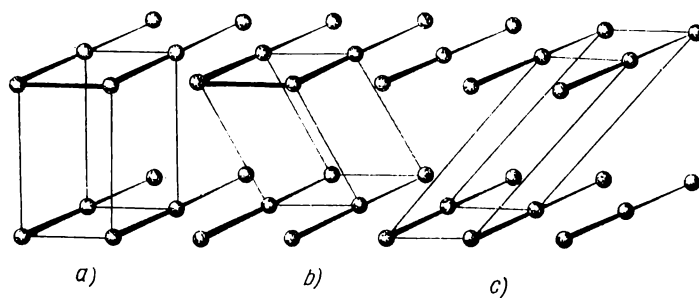


Fig. 263

In a number of cases, it is expedient to select an elementary cell so that its volume is greater than that of a primitive cell. Thus, in order to take maximum advantage of the symmetry of a crystal, we often select an elementary cell with an additional lattice point at face centres or at the centre of the cell. Three cases are encountered frequently:

(1) Body-centred cell. An additional lattice point is located at the intersection of the spatial diagonals of the cell. In this case, there are two lattice points per cell:  $[[000]]$  and  $\left[\left[\frac{1}{2} \frac{1}{2} \frac{1}{2}\right]\right]$ . The lattice point at the centre of a cell belongs entirely to the given cell. The eight lattice points at the corners are shared jointly by eight cells, i.e.,  $\frac{1}{8}$  of each of these lattice points belongs to the given cell.

(2) Face-centred cell. An additional lattice point is located at the centres of a pair of faces, e.g.,  $ab$ . In this case, too, there are two lattice points per cell:  $[[000]]$  and  $\left[\left[\frac{1}{2} \frac{1}{2} 0\right]\right]$ .

(3) All-sided face-centred cell. Additional lattice points are located at all face centres. In this case, there are four lattice points per cell:  $[[000]]$ ,  $\left[\left[0 \frac{1}{2} \frac{1}{2}\right]\right]$ ,  $\left[\left[\frac{1}{2} 0 \frac{1}{2}\right]\right]$  and  $\left[\left[\frac{1}{2} \frac{1}{2} 0\right]\right]$ .

The following designations are commonly used:  $P$ —primitive cell;  $A$ ,  $B$  and  $C$ —face-centred cells with a lattice point in faces  $bc$ ,  $ac$  and  $ab$ , respectively;  $F$ —all-sided face-centred cell, and  $I$ —body-centred cell.

As emphasised earlier, a lattice point is an arbitrary point, but for convenience is selected in a specific manner. The succeeding lattice point is separated from the selected one by a distance equal to the lattice spacing. Thus, there is one lattice point per primitive cell.

All the atoms of a primitive cell may be replaced by lattice points. Usually, a point of intersection of symmetry elements is taken as a lattice point.

A primitive cell may consist of many atoms. When there are many atoms in a cell, the structure is described by the coordinates of the atoms in an elementary cell.

The model of a crystal as a space lattice is in full accord with experimental data. Crystal edges and faces correspond to nodal straight lines and planes. The angles between crystal faces and edges are the same for all crystalline objects of a given chemical compound.

The symmetrical features of crystal structure may also be determined from the space lattice model.

Crystals of different substances have different symmetry. If a crystal is well formed, its symmetry is self-evident. It is clear that the planes and axes of symmetry may be passed through a crystal in a specific manner.

The external symmetry of a crystal can be explained by its internal structure, i.e., the symmetry of the space lattice.

In addition to axes of symmetry, symmetry elements encountered in crystals include mirror planes and inversion or symmetry centres. Fig. 264 illustrates the operations which may be performed by means of these symmetry elements.

It has been long known that axes of symmetry of fifth order and axes of symmetry of higher order than the sixth do not occur in crystals\*. Basing himself on this fact, A. V. Gadolin proved that there can be only 32 groups of symmetry among crystals.

Before the development of the space lattice theory, it was unclear why axes of fifth, seventh, etc., orders were not encountered in crystals. This and other features of crystal symmetry can be explained by the space lattice theory.

\* If a body is turned about a certain axis by an angle  $2\pi/n$  so that the figure obtained coincides with the original figure, such an axis is called an axis of symmetry of  $n$ -th order. For example, a 3-faced prism based on an isosceles triangle has an axis of symmetry of third order which passes through the centre of the triangle and is parallel to the edges of the prism.

Let us consider rotation of a lattice plane. Rotations which are not possible for a plane will certainly not be possible for the entire lattice.

Assume that an axis of  $n$ -th order passes through the lattice point  $B$  and that the identical one closest to it passes through the lattice point  $A$  (see Fig. 265). Rotation about axis  $B$  transfers lattice point  $A$  to  $A'$  and rotation about axis  $A$  transfers lattice point  $B$  to  $B'$ . It is seen from the diagram that  $B'A' = AB \times \angle (1 + 2\cos \alpha)$ . But the distance  $A'B'$  must be a multiple of the lattice spacing  $AB$ , for  $A'B'$  is parallel to  $AB$ . Therefore,  $2\cos \alpha$  must be equal to a whole number. It follows that  $\cos \alpha$  can assume only the values  $0, \pm\frac{1}{2}$  and  $\pm 1$ , and  $\alpha$  the values

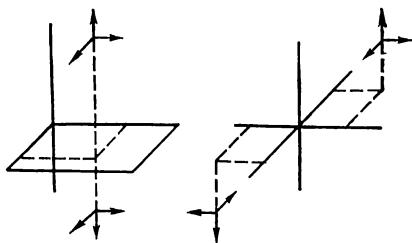


Fig. 264

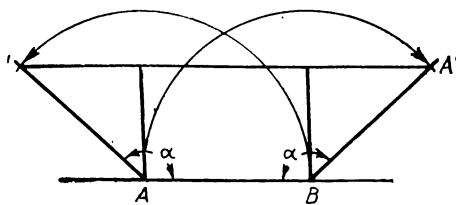


Fig. 265

$60^\circ, 90^\circ, 120^\circ, 180^\circ$  and  $360^\circ$ . This also follows from the definition of a closed symmetric operation: the rotation angles are equal to  $360^\circ$  divided by a whole number. Thus, it is possible to have rotational axes of sixth, fourth, third, second and first orders in a crystal.

One can prove easily that an axis of symmetry is a nodal line which is normal to a nodal plane.

The symmetry of a crystal is determined by the symmetry of the space lattice. But it should be realised that the symmetry of a lattice is considerably richer.

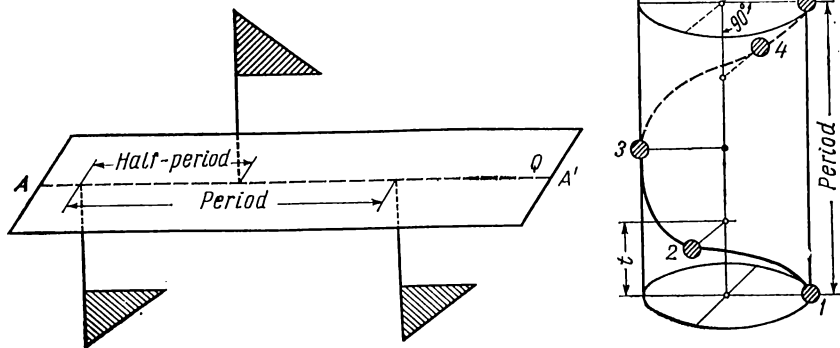


Fig. 266

Only 32 groups of crystal symmetry exist but, as was first shown by E. S. Fyodorov, the founder of structural crystallography, a space lattice has 230 types of symmetry (Fyodorov groups).

The richer symmetry of a lattice is due to the fact that in addition to closed symmetric operations it includes a symmetry element which is not possible for a body,

i.e., translation. A symmetric operation consists in the displacement of a body to a position which is indistinguishable from its original position. Therefore, in the case of an infinite lattice, a displacement along one or another nodal line is a symmetric operation.

Translation introduces the following new symmetry elements: (1) a combination of rotation and translation—screw axes; and (2) a combination of reflection and translation—slip planes (Fig. 266). A screw axis of fourth order is shown in Fig. 266 (right diagram). Each of points 2, 3, 4 and 5 is obtained from the preceding one by a  $90^\circ$  turn and displacement along the axis by  $\frac{1}{4}$  of a period ( $t$ ). In Fig. 266 (left diagram) we see triangles reflected in a plane  $Q$  and slipped along  $AA'$  by  $\frac{1}{2}$  of a period.

#### Sec. 240. THE PACKING OF PARTICLES IN A CRYSTAL

Figures having a definite volume and shape can be stacked or packed. It is by no means clear to what extent a formation of atoms will conform to this picture. Here, the answer to the following question is of prime importance: if we attribute

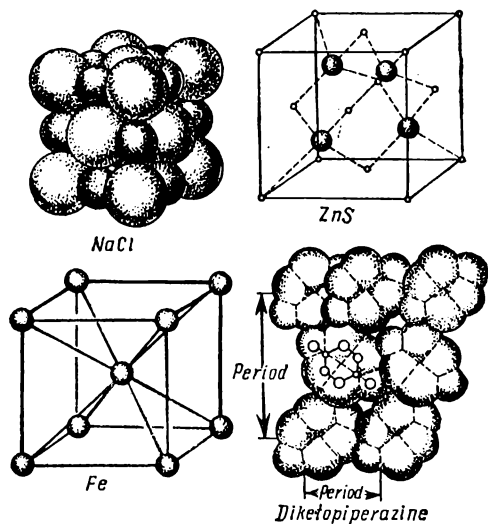


Fig. 267

a definite shape to an atom or group of atoms, will this even roughly correspond to a minimum on the potential curve of particle interaction in a given direction? Moreover, to what extent will the volume attributed to the atom or group of atoms encompass all the electrons, including the valence electrons, belonging to this atom or group of atoms? If it turns out that describing the limits of an atom or molecule by a definite contour has physical meaning, we will have ascertained at the same time how this shape is manifested when a crystal is formed from particles.

The nature of interactions between atoms in a crystal is infinitely varied. However, several limiting cases can be singled out: pure ionic bonds, homopolar bonds, metallic bonds and molecular bonds. As examples, let us consider the structures

of rock salt, zinc sulphide, iron and the organic substance diketopiperazine (see Fig. 267). In the first three cases, the absence of a tight group of atoms is characteristic. A molecule cannot be distinguished in a crystal of rock salt. Each atom of sodium has six perfectly equal chlorine atoms as neighbours. Nor can a molecule be distinguished in the other two compounds, which are examples of homopolar and metallic bonds.

In ionic crystals, the particles consist of positive and negative ions, which attract one another in accordance with the laws of electrostatics. Everything said about ionic bonds in molecules applies to crystals. The nature of ionic structures is reasonably well if the ions are represented by spheres having definite ionic radii. The values of ionic radii in ionic molecules differ little from those in crystals.



In crystals with a homopolar bond, every two atoms are linked by a pair of electrons. In this manner each zinc atom is linked with its surrounding sulphur atoms, and each sulphur atom with its surrounding zinc atoms. If a sign of a homopolar bond were considered to be an indication of a molecule, it could be said that an entire crystal constituted a single molecule. It is physically meaningless to consider homopolar crystals as constructed of contiguous spheres. The dimensions of free atoms of sulphur and zinc are considerably greater than the distance between them in zinc sulphide. A homopolar bond brings the atoms into close contact and makes regions in which electrons of these atoms are located common regions. If the structure of zinc sulphide consisted of contiguous spheres, a large portion of the electron cloud would be located outside the spheres, i.e., only 25 per cent of the volume would be occupied by spheres.

Metallic bonds will be discussed in Chapter 37. Here, a few words on this subject will suffice. In metals and alloys, outer electrons are common, forming an electron "gas". The lattice of a metallic compound consists of atomic residues (positive ions) "cemented" by electrons. Here, too, it is physically meaningless to represent the structure as contiguous spheres, in spite of the fact that formally the structures of certain metals can be represented as closely packed spheres.

In crystals of the same type as diketopiperazine the molecules are distinct. They can be easily recognised since the intermolecular distances are considerably greater than the intramolecular distances. By studying the arrangement of molecules in crystals, crystallographers have been able to pick out intermolecular radii of portions of spherical surfaces describing the limits of a molecule. The model of a molecule formed of microscopic spheres of intermolecular radius is based on an analysis of crystal structures. Such a geometric representation of crystals formed of molecules is quite justified since most of a molecule's electron cloud is contained within the contour of the molecule.

It must be concluded that the representation of the component particles of a crystal as geometric figures is meaningful in two cases: in ionic and in molecular crystals. On the other hand, such a representation is meaningless in the case of homopolar crystals and metallic compounds.

Now, the following question arises: how is the shape of ions and molecules manifested in the formation of a crystal? The answer is that it is manifested in the compact packing of particles. Experiments indicate that molecules are always packed in such a way that a "projection" of one molecule fits into a "depression" of another. There is a clear tendency for the molecules to become so oriented with respect to one another that the volume of an elementary cell is as small as possible. The situation is similar in the case of ionic crystals. The stacking of spheres occurs in such a manner that the large spheres fit closely together, while the small spheres (ions) fit into the empty spaces of the basic structure.

In representing ions by spheres, and molecules by spatial figures, we find that the "empty" space is equal to 25-35 per cent of the total.

Close packing in molecular and ionic crystals provides basic proof that shape and volume are meaningful attributes of atoms and molecules.

#### Sec. 241. MOLECULAR CRYSTALS

The assertion that molecules are bound by intermolecular forces in a molecular crystal is, of course, pure tautology. What then can be said about this concept and about the nature of intermolecular forces?

One should distinguish between polar and nonpolar molecules. Polar molecules have a pronounced dipole moment which appears due to the fact that electron den-

sity turns out to be shifted to one side. Nonpolar molecules may be regarded as the totality of neutral atoms, whereas polar molecules may be thought of as comparatively small charges on atoms superposed on a nonpolar framework of a molecule. According to such a conception, in the general case, interaction of molecules is successively described as interaction of neutral particles plus electrostatic action.

Experiments and calculations show that interaction of electrically neutral particles constitute the bulk of the energy of interaction of molecules. This follows, for instance, from the fact that the values of the heat of sublimation (which is a good measure of interaction of molecules) for nonpolar benzene and polar nitrobenzene are close to each other.

Thus, all molecules should be considered as systems of electrically neutral atoms. The energy of interaction of molecules may be represented (with sufficiently good approximation) as a sum of energies of interaction of atoms (the additivity principle).

The curve of the potential energy of interaction of two atoms belonging to different molecules is represented, of course, by a general-type curve (both for valence-bonded atoms and atomic nuclei) having a steep rise in the direction of small distances, a flat rise with an asymptotic approximation to zero in the direction of great distances, and a minimum (a potential well) for a certain equilibrium distance. The abscissa of this minimum is just the intermolecular radius by means of which the shape of a molecule is framed.

While the abscissa of the well bottom lies for valence-bonded atoms at 1 to 1.5 Å, and the well depth is measured by many tens of J/mol, the corresponding figures for atoms of different molecules will be 2-4 Å and tenths of J/mol, respectively.

Neutral atoms are all the same electrical systems. Therefore, we must try to explain the shape of the curve of non-valent interaction in terms of electronic structure of atom.

First of all, what about the origin of attraction? A quantum-mechanical explanation of these forces which are known as *dispersion forces* was given by H. London. (London's equation for intermolecular attraction.) Here is the explanation itself:

All molecules possess energy even at 0°K (null point energy) and this fact can only be explained by assuming that, even at this temperature, the nuclei and electrons vibrate in some way with respect to each other. As an instantaneous picture, the molecules will show various arrangements of the nuclei and electrons, these arrangements giving rise to dipole moments. A summation of these dipole moments over all the molecules will give a zero resultant. The cohesive force between molecules was attributed by London to the transient dipoles induced in molecules in phase with themselves by these temporary dipoles. The interaction energy has been calculated to a first approximation as

$$U_D = -\frac{3}{4} \frac{h\nu_0\alpha^2}{r^6}$$

where  $\nu_0$  is the characteristic frequency of the molecules,  $\alpha$  is equal to  $e^2/k$  where  $e$  is the electron charge,  $k$  is the restoring force constant, and  $r$  is the equilibrium distance between the positive ends of the dipoles.

The interaction force is expressed, if it is needed, by the following general formula

$$F = -\frac{\partial U}{\partial r}.$$

The attraction forces start noticeably acting at short distances. Mutual overlapping of electronic shells (squares of wave functions) is obstructed by repulsion of electrons. But the leading role is played not by the electrostatic force acting

between like charges, but by the Pauli principle which states that it is impossible for a third electron to approach a region occupied by a pair of electrons with opposite spins.

A geometrical model of the molecule simplifies the picture of interaction. Speaking of a molecule having a rigid form, we, as a matter of fact, replace the described potential by a rectilinear one, as is shown in Fig. 268. For a number of purposes, such simplification is completely justified. It turns out that the shape of a molecule determines not only the intermolecular distances, but also the character of the molecular packing. Molecular polarity and other peculiarities of the forces of attraction are not only incapable of affecting the intermolecular distances, but also fail to disturb the tendency to compact packing of molecules. Thus, in practically all cases, minimum energy is achieved by compact packing of molecules. A strict subordination of molecular crystals to the principle of compact packing leads to a scarce variety of structural types and symmetry is encountered among such crystals.

It is convenient to view the packing of molecules in crystals as a tight packing of compact layers. Two types of such layers are encountered. These are illustrated in Fig. 269. The one with right-angled cells has greater symmetry. In this case (more than 90 per cent of organic crystals are formed of such layers), the molecules are stacked in the characteristic zig-zag manner shown in the figure. The rows of molecules forming a layer are connected by a screw axis of the second order ( $2_1$ ).

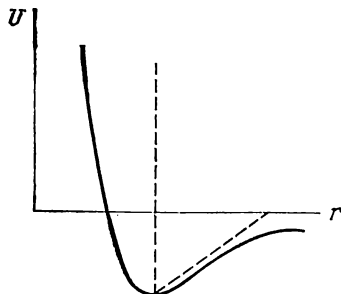


Fig. 268

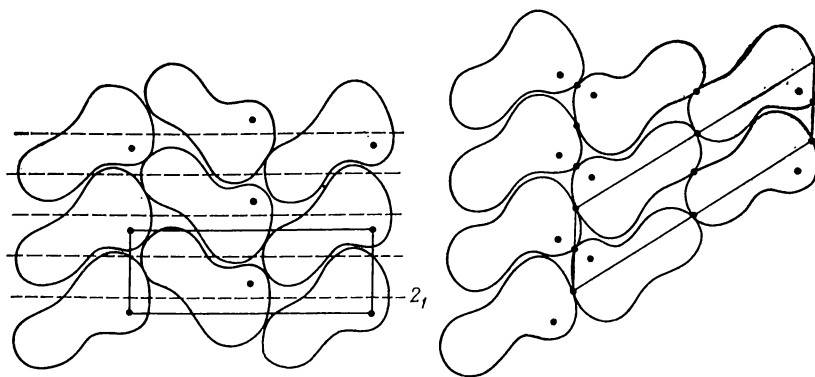


Fig. 269

This means that one row of molecules can be transformed into an adjacent one by a  $180^\circ$  turn and a displacement of half a period along the axis.

In compact layers, each molecule has six close neighbours. When the layers are stacked, a molecule usually obtains six additional close neighbours—3 above and 3 below. Thus, the total number of such neighbours becomes equal to 12.

Crystals having a high order of symmetry are rarely encountered in the world of molecular crystals. It is not possible to pack compactly unsymmetrical molecules in symmetrical crystals.

If a molecule possesses symmetry, this does not mean that the crystal also has such symmetry. A molecule of naphthalene has a high order of symmetry; three

mutually perpendicular planes of mirror symmetry may be drawn through it (see Fig. 270). If these symmetry elements were preserved in a formation of molecules, the packing would be insufficiently compact. Therefore, the symmetry elements of a molecule which prevent greater compactness are "lost" in the formation of a crystal. The preservation of a centre of inversion is possible without sacrificing com-

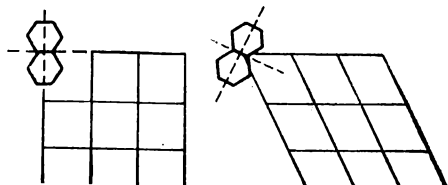


Fig. 270

compactness of molecular packing. A crystal formed of molecules possessing this symmetry element usually does not preserve other symmetry elements of the molecules, but does preserve a centre of inversion.

In other cases as well, the outcome of the tendency to symmetry as opposed to the tendency to compactness can be reliably predicted.

An example of a packing arrangement typical of molecular crystals is provided by diketopiperazine (see Fig. 267). The molecules have a high order of symmetry, but the crystal preserves only a centre of inversion. A molecule, of course, does not cease to be highly symmetrical simply because a crystal of such molecules does not possess its symmetry elements.

#### Sec. 242. COMPACT PACKING OF SPHERES

A very important class of ionic crystals may be represented by a compact packing of spheres.

Most anions are larger than cations. In such cases, crystals constitute a compact packing of anion spheres between which cations are located. This is how silicates, one of the largest groups of natural inorganic substances, are formed. In silicates, the cations are located in the empty spaces of a compact structure of oxygen anions.

Let us examine the laws of compact packing of spheres, i.e., the fundamental structures of a great number of crystals. The only possible arrangement of a compact layer of spheres is shown in Fig. 271. Each sphere has six neighbours. To form a compact packing arrangement, one must place the spheres of a second layer in the spaces of the layer below it. It is not possible to fill all the spaces with spheres of the same size: every second space in the figure is filled (crosses denote the spaces of the first layer which are filled by spheres of the second, and dots denote the spaces which remain vacant).

There is also only one possible arrangement for compact packing of two layers of spheres.

For the third layer, however, the situation is different. In order to achieve compact packing in this case, one must place the spheres of the third layer in the spaces of the second, but this can be done in two ways: the centres of the spheres of the third layer may be placed either above the centres of the spheres of the first layer or above the spaces denoted by dots. Both three-layer structures are packed equally compactly; however, they differ significantly from one another. When a fourth layer is added to the structure, the number of possible packing arrangements

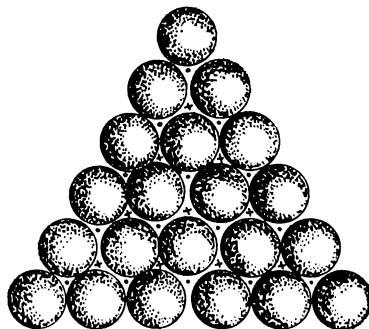


Fig. 271

becomes even greater: from the two three-layer structures, one can make four four-layer structures. In the case of a five-layer structure, there are five possible arrangements, etc. It is evident that the number of different arrangements of spheres packed equally compactly increases very rapidly with increasing number of layers.

Now, let us compare a crystal lattice with such an arrangement of spheres. A crystal may be represented as a structure of spherical atoms in which the arrangement of layers is repeated exactly after a certain number of layers. If such a

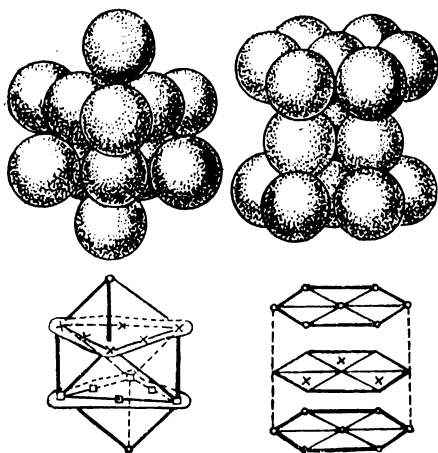


Fig. 272

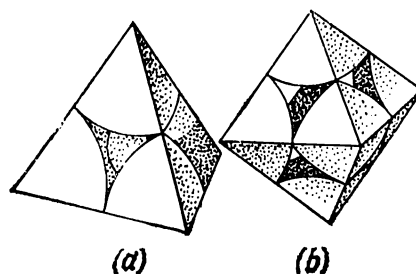


Fig. 273

sequence begins with the fourteenth layer, for example, this means that a cell is thirteen layers high. In such a case, the fourteenth layer is above the first, the fifteenth above the second, the sixteenth above the third, etc.

The simplest packing arrangement consists of two layers: the third layer lies above the first, the fourth layer above the second, etc. (see Fig. 272, right). This is a hexagonal compact packing arrangement. A cell of such a crystal is shown in the lower right-hand corner of the figure. The dots and crosses denote the locations of the centres of the spheres.

Three-layer crystals, in which the fourth layer is a repetition of the first, the fifth of the second, etc., are very common (see Fig. 272, left). In the lower left-hand corner of the figure, where only the centres of atoms are indicated, we see that a cubic elementary cell, centred in all faces, may be selected. Here, the compact layers are arranged perpendicular to the spatial diagonal of a cube. Such a structure is called a cubic compact-packing arrangement.

Empty spaces remain in a packing of spheres of equal size. It can be easily calculated that the volume of such spaces is equal to about  $1/4$  of the overall volume. There are two kinds of such empty spaces: one is surrounded by four spheres the centres of which are located at the vertexes of a regular tetrahedron (see Fig. 273a); the other is surrounded by six spheres—the centres of these spheres form a regular octahedron (see Fig. 273b). The first kind is smaller in size, but there are twice as many of them as the second.

It can be shown that in any compact arrangement of equal spheres there are two small empty spaces and one large one per sphere. Small spheres can fit into these spaces but if they are somewhat too large, they cause the large neighbouring spheres to move apart, loosening the compact packing arrangement.

Since different packing arrangements are possible with equal numbers of spheres, and small spheres may fill the empty spaces in different ways, ionic crystals have a great variety of structures.

In crystals of common salt, a compact three-layer structure is formed by large chlorine ions (light spheres in Fig. 267), and sodium ions (dark spheres) fill all the large spaces; hence, every sodium atom is surrounded by six chlorine ions. In iron disulphide (pyrite), a compact two-layer structure is formed by large sulphur ions; iron ions fill all the large spaces. In a crystal of lithium oxide, the chemical formula of which shows that there are two lithium atoms for every oxygen atom, the compact structure is formed by large oxygen ions. Since lithium ions fill all the small spaces, each lithium ion has four neighbours (oxygen ions). In a crystal of cadmium chloride, the chemical formula of which shows that there are two chlorine atoms for every cadmium atom, the compact structure is formed by large chlorine ions; cadmium ions fill large spaces but not all, i.e., they fill the large spaces of every third layer of chlorine ions. We have presented, of course, only the simplest "patterns" of the filling of empty spaces in compact packing arrangements.

#### Sec. 243. EXAMPLES OF CRYSTAL STRUCTURES

The largest group of crystals consists of bodies formed of molecules. Ionic compounds also constitute a fairly large group. As indicated already, the representation of a crystal in these cases as closely packed particles is entirely justified. However, it is necessary to examine those structures in which the direction of the bonds between atoms, the deviation of the electron cloud from spherical symmetry, etc., are the cause of structural arrangements which cannot be represented so simply. Such exceptions include structures of atoms bound by common electrons.

Most metals have structures consisting of body-centred cubic cells. In such crystals, each atom has eight neighbours rather than twelve, as is the case in a compact packing of spheres. This is the case, for example, for atoms of iron (see Fig. 267). The lattice of iron is cubic; iron atoms are located at the corners and centres of cubes. Lithium, potassium, caesium and a number of other substances possess such a structure.

In Fig. 274, the structure of crystalline mercury is compared with an ideally cubic compact packing arrangement. It can be seen that the nature of the arrangement of atom centres is the same in both cases, but in the former the distance between layers is less, and the distance between atoms in a layer is more than in the latter. This is analogous to a compact packing of slightly flattened spheres.

Many examples exist of such more or less "damaged" compact packing structures. In the case of ice (see Fig. 275), all resemblance to a spherical packing arrangement is lost. The link between each pair of oxygen atoms is implemented by a hydrogen atom. In these four bonds, there are two oxygen atoms per hydrogen atom. (The structure shown in Fig. 275 does not, of course, contradict the chemical formula for water.) For purposes of clarity, a "hydrogen" bond is shown in the figure as a "neck". The structure of ice is very loose, as indicated by the large "holes". If one projects the structure above the plane of the figure, these holes are transformed into broad channels which pass through the structure. The structure of ice is an important exception to the general rule. This does not mean that the cases in which the likening of a crystal to a compact packing of particles is meaningless are rare.

As indicated above, crystals formed of atoms bound by common electrons cannot be likened to a compact packing of spheres.

The structure of zinc sulphide, which is illustrated in Fig. 267, is quite typical. Moreover, several elements reveal a similar structure. These include carbon (diamond), silicon, germanium and tin (white).

Homopolar bonds can form layers and chains of atoms.

Fig. 276 shows the structure of graphite. The carbon atoms in graphite form a layered structure, but these layers are not the same as those of a compact packing

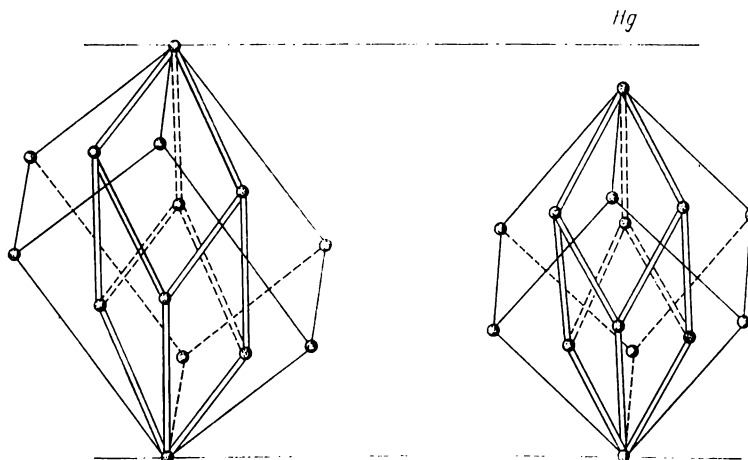


Fig. 274

arrangement. It is not possible to form a layer of graphite of contiguous spheres. In graphites, layers of strongly bound atoms constitute planes. Arsenic and phosphorus also form layered structures in this sense, but the atoms of their layers are not arranged in a single plane. An example of a structure consisting of chains of

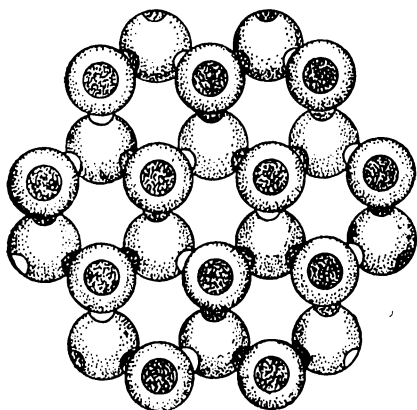


Fig. 275

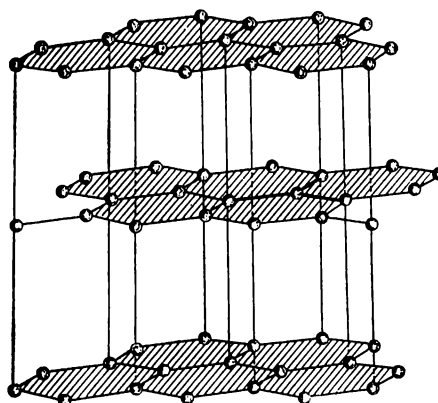


Fig. 276

strongly bound atoms is provided by gray selenium. Each atom of this substance is strongly bound to only two neighbours. In gray selenium, the atoms form an endless spiral about a straight line. The separation between atoms of neighbouring spirals is considerably greater than the separation between close atoms in a given spiral.

The black, lustreless, soft graphite used in pencils and the shiny, transparent, hard diamond which can cut glass are composed of the same kind of atoms, i.e., carbon. This is a very striking example of how greatly the properties of a crystal are affected by the arrangement of its atoms. Refractory crucibles capable of withstanding temperatures of 2,000-3,000°C are made of graphite, while at temperatures exceeding 700°C diamonds are burned up; graphite has a specific weight of 2.1 as compared with 3.5 for diamond; graphite conducts electric current, but diamond does not; etc.

The ability to form different crystals is not a property of carbon alone. Almost every chemical element in the crystal state and every substance has several forms. Six forms of ice are known, nine of sulphur, and four of iron.

At room temperature, atoms of iron form cubic lattices, the atoms being located at the corners and centres of cubes. Thus, each atom has eight neighbours. At high temperatures, iron atoms form a compact structure. Here, each atom has twelve neighbours. Iron possessing eight neighbours is soft, while iron possessing twelve is hard. The quenching of steel fixes, at room temperature, a compact cubic structure which is stable at higher temperatures.

We have seen in the cases of carbon and iron that the structures of different crystals of one and the same substance may differ considerably from one another. The same holds true for other substances.

Thus, for example, in a crystal, yellow sulphur forms corrugated rings of eight atoms each. Each ring constitutes a sulphur molecule. Red sulphur also consists of such rings but they are turned completely differently with respect to one another.

Yellow phosphorus atoms form cubic structures with eight close neighbours. Black phosphorus has a layered structure similar to graphite.

Gray tin has a structure similar to that of a diamond. Theoretically, white tin could be obtained from gray tin by strongly compressing the diamond-like structure along the axis of a cube. As a result of this flattening process, a tin atom would have six close neighbours instead of four.

Organic substances also frequently have a variety of crystal forms. The very same molecules are arranged differently with respect to one another.

#### Sec. 244. THERMAL VIBRATIONS IN A CRYSTAL

From the viewpoint of energy, an ideal crystal is, in a way, the antithesis of an ideal gas.

In an ideal gas, the interaction energy of particles is much less than the average energy of thermal motion  $kT$ . On the other hand, since strong coupling exists between particles in a crystal, the interaction energy is much greater than  $kT$ . Therefore, thermal motion in crystals cannot disrupt the coupling between atoms, but merely results in small vibrations of the atoms about equilibrium positions.

In a crystal, every atom vibrates about an equilibrium position. For most crystals, the vibration amplitudes are of the order of  $0.1 \text{ \AA}$ , i.e., a small fraction of the distance between close atoms, which, as we know, is of the order of  $1.5\text{--}2 \text{ \AA}$ .

The nature of this vibration may be very complex. During a vibration period, an atom describes a complex curve about its equilibrium position. This is due to the fact that an atom is bound to its neighbours by different forces; hence, its vibrations are of an anisotropic nature. In any case, it is always possible to resolve the vibrations of an atom along three axes. Evidently, the atoms of a crystal will have  $3N$  degrees of freedom, where  $N$  is the number of atoms.

If the molecules of a crystal are clearly distinguishable, it is meaningful to speak of vibrations of a molecule and vibrations of atoms within a molecule. Since the



coupling between molecules is considerably weaker than between atoms, the frequencies of their vibrations will be less. In molecular crystals, the motion of a molecule as a whole is of decisive importance. The vibrations of a molecule about its equilibrium position are of a translational as well as torsional nature. Apparently, it is even possible in rare cases for total rotation of molecules about centre of gravity to occur. For example, such rotation of molecules probably occurs in the case of solid methane ( $\text{CH}_4$ ).

The total energy of a vibrating particle consists of its potential energy and its kinetic energy. The average values of these two energies over a period of vibration are equal to each other. As is known, the average kinetic energy of an atom in a

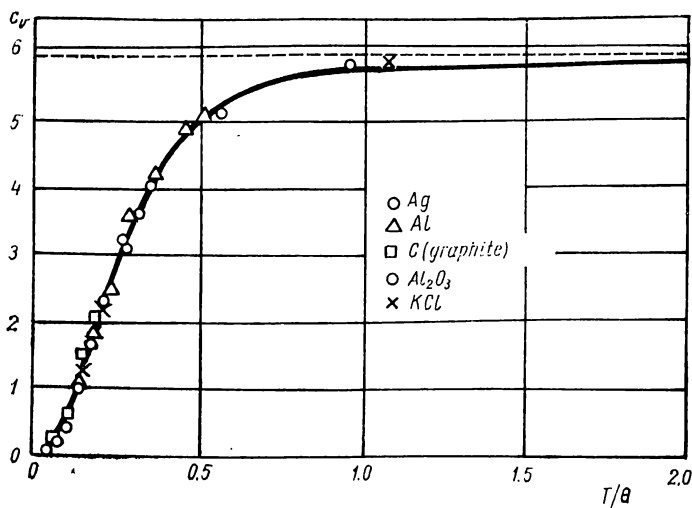


Fig. 277

gas is equal to  $\frac{3}{2} kT$ . It would seem reasonable to assume that, at the same temperature, a vibrating atom possessing twice the average energy possesses  $3 kT$  units of thermal energy. Then, one mole of crystalline substance should have an energy  $3RT$  and a molar heat capacity  $c_v = 3R \approx 6 \text{ cal/mole} \approx 24.93 \text{ J/mole}$ .

At high temperatures, this formula is in very close agreement with experimental results. The temperature dependence of the heat capacity of crystalline bodies is shown in Fig. 277. Beginning at zero, the heat capacity increases, attains a value of 6 cal/mole at a certain temperature, and then remains unchanged. The ratio of temperature to a constant  $\theta$  (to be discussed in the following article) is plotted along the x-axis.

At high temperatures, the value of  $kT$  is significantly greater than the difference between vibrational energy levels; hence, the quantum nature of distribution of vibrating atoms according to energy levels does not affect the value of the average vibrational energy. Under such conditions, the simplified method of calculating the average energy is quite justified. This is confirmed by exact calculations which take into account the distribution of atoms according to energy as given by the Boltzmann law.

When  $kT$  becomes comparable to the difference between energy levels, the Boltzmann law is no longer applicable and must be replaced by a quantum distribution law (see p. 541). Calculations indicate that heat capacity decreases with decreas-

ing temperature. We shall not present these calculations; we note, however, that qualitatively a decrease in  $c_v$  with decreasing temperature is quite understandable. The smaller the magnitude of  $kT$ , the smaller the number of energy transitions that may occur in a system. This means that the possibility of thermal exchange decreases with decreasing  $kT$  and approaches zero as a limit. The limiting case can be explained as follows: energy cannot be transmitted to a body in extremely small bursts of  $kT$  even if there are an infinite number of such "bursts" with a very high total energy. Energy cannot be transmitted since a single "burst" does not suffice to transfer a system from its zero-energy level to the next one.

Experiments indicate that energy transitions corresponding to a change in the state of molecular motion in a crystal lie in the region of long infrared waves.

For purposes of guidance, let us assume that such transitions correspond to a wavelength of 1 mm.

Let us compare several values of  $kT$  with the quantum energy corresponding to a wavelength of 1 mm. For this wavelength,  $\nu = 3 \times 10^{11} \text{ sec}^{-1}$ , i.e.,  $h\nu = 200 \times 10^{-17} \text{ erg}$ .

Temperature (K)	$kT$ (ergs) (Approx. values)
500	$7,000 \times 10^{-17}$
100	$1,380 \times 10^{-17}$
10	$138 \times 10^{-17}$
1	$14 \times 10^{-17}$

It can be seen that at 100 K the thermal vibration energy still considerably exceeds the difference between the energy levels of a molecule in a crystal. At 10 K it has the same order of magnitude and at 1 K thermal vibrations are unable to produce transitions from one level to another.

#### Sec. 245. THERMAL WAVES

An interesting feature of thermal vibrations in a crystal is their occurrence in the form of thermal waves. Therefore, atomic vibrations cannot occur independently of one another. An atom deviating from its equilibrium position pulls along the next one.

Since a crystal is a finite body, standing waves are formed within it. As in the case of all natural oscillations, the maximum length of a standing wave equals twice the dimension of the body. The boundaries of a crystal must correspond to standing-wave nodes.

On p. 101 we discussed elastic vibrations of solids considered as a continuum. It was shown that there arise in a finite solid body numerous standing waves of different direction and frequency. The picture becomes considerably more complex when the atomic structure of the solid is taken into account. A theoretical investigation of the possible vibrational motion of atoms in a monocrystal indicates that thermal motion in a crystal can be represented as the result of the superposition of  $3sN$  waves, where  $N$  is the number of cells and  $s$  the number of atoms per cell. The number of possible waves is equal to the number of degrees of freedom of the system of atoms forming the crystal. How do these waves arise and how are they manifested?

Let us restrict ourselves to a consideration of a chain of atoms, i.e., a "unidimensional crystal". Fig. 278 shows such an atomic chain, the "cells" of which consist of two atoms, denoted by black and white dots. Since the actual thermal motion of atoms in a crystal is of a very complex nature, the figure has been simplified

to show the "elementary" waves into which this motion can be resolved. Calculations indicate that the resulting vibration can always be represented as the sum of harmonic vibrations. Like in the case of a solid rod, a series of waves of various wavelengths arise in a unidimensional crystal. If a chain consists of a thousand cells with period  $a$ , there arise  $N$  waves of wavelength  $2,000 a$ ,  $1,000 a$ ,  $\frac{2,000}{3} a$ ,  $500 a$ ,  $400 a$ , etc. The shortest wavelength is  $2 a$ .

But this is not all. Each of the possible wavelengths occurs in  $s$  variations. Two types of waves of the same wavelength are shown in the figure. A case exists in which the lattice of atoms vibrates as a whole. Such a wave is called an *acoustical* wave. The remaining  $s - 1$  waves are quite different. In these cases, different types of atoms execute complex motion with respect to one another and at each instant only atoms of a single type fall on the sinusoid. There are  $s - 1$  such vibrations, which are called *optical* vibrations.

The figure shows waves corresponding to atomic vibrations in a single direction.

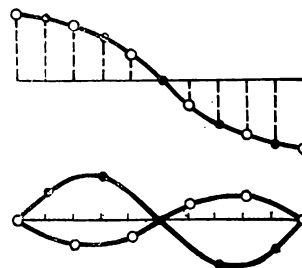


Fig. 278

Atomic vibrations can always be resolved into two transverse components and one longitudinal component. Therefore, a wave of given wavelength travelling in a given direction will have 3 components of the acoustical type and  $3(s - 1)$  of the optical type. Of the total of  $3sN$  waves,  $3N$ —two transverse and one longitudinal for each wavelength in each direction—will be acoustical. This is completely valid for three-dimensional crystals.

Although the wavelengths and frequencies of the waves have discrete values, we may utilise the results obtained on p. 101 and express, approximately, the number of acoustical vibrations having frequencies less than  $\nu$  as

$$\frac{4\pi\nu}{3c^3} \int_0^{\nu} \nu^3 d\nu.$$

Here,  $\nu$  is the volume of the crystal and  $c$  the velocity of the wave. The velocities of the longitudinal and transverse waves are different. Therefore, the total number of waves, which is equal to  $3N$ , should be written as follows:

$$\frac{4\pi\nu}{3} \left( \frac{1}{c_l^3} + \frac{2}{c_t^3} \right) \nu_{\max}^3 = 3N,$$

where  $c_l$  is the longitudinal velocity and  $c_t$  the transverse velocity of the wave. Whence, the value of the maximum frequency of vibration,  $\nu_{\max}$ , can be easily found. The corresponding wavelength,

$$\lambda_{\min} = \frac{c_l}{\nu_{\max}} = \frac{c_t}{\nu_{\max}},$$

is, as it should be, of the same order of magnitude as a cell period.

If the velocity of propagation of acoustical waves in a crystal is known, we can calculate  $\nu_{\max}$ , the value of which determines to a large extent the behaviour of a crystal.

As was indicated above, heat capacity depends on the commensurability of the energy  $h\nu$  and the thermal energy  $kT$ . If  $h\nu_{\max} \ll kT$ , thermal exchange excites all of the vibrations and waves in a crystal, i.e., all of the quantum transitions are possible. As a result, the quantum nature of thermal exchange is not apparent.

Such a crystal has a characteristic temperature  $\theta = \frac{h\nu_{\max}}{k}$ , which is much less than the temperature of the experiment, i.e.,  $\theta \ll T$ . On the other hand, if  $h\nu_{\max} \gg kT$ , i.e., if the characteristic temperature  $\theta \gg T$ , only vibrations of low frequency are excited in the crystal because high energy levels cannot be surmounted by the thermal "bursts".

Here are examples of the characteristic temperatures<sup>\*</sup>(K) of a number of crystals:

Pb	Benzene	Ag	NaCl	Be	Fe	Diamond
90	150	215	280	1,000	450	1,860

For such substances as lead and benzene, room temperature is "high". This corresponds to the horizontal portion of the heat capacity curve ( $c_v = 6$  cal/mole; see Fig. 277). On the other hand, room temperature is low for beryllium and diamonds. Thermal vibrations of these substances are excited to an insignificant extent and their heat capacity is considerably less than 6 cal/mole. The maximum vibration frequencies  $\nu_{\max} = \frac{k\theta}{h}$  can be calculated from the above values of the characteristic temperature  $\theta$ :

Pb	Benzene	Ag	NaCl	Be	Fe	Diamond
$1.88 \times 10^{12}$	$3.13 \times 10^{12}$	$4.47 \times 10^{12}$	$5.84 \times 10^{12}$	$20.8 \times 10^{12}$	$9.3 \times 10^{12}$	$38.8 \times 10^{12}$

It can be seen that the limiting frequencies of thermal vibrations lie, as was assumed in the example on p. 480, at the boundary between the infrared and radio bands ( $\lambda_{\min} > 10^{-2}$  cm).

#### Sec. 246. THERMAL EXPANSION

How can we explain the fact that the average distance between neighbouring atoms increases with increasing temperature? To answer this question, let us consider the curve of potential energy of interaction between atoms or molecules (see Fig. 279). Irrespective of the peculiarities of the interaction between particles, the potential curve is always asymmetrical: in the direction of decreasing separation between particles the curve rises steeply, while in the direction of increasing separation a well wall is formed. This is due to the following simple fact: practically, the separation between two atoms or molecules cannot be decreased indefinitely, but it can be increased indefinitely—at great distances the bond between the particles is broken.

The maximum and minimum distance between vibrating atoms may be noted on a potential curve. The middle of the segment connecting the two limits corresponds to the average position of an atom. When the temperature increases from  $T_1$  to  $T_2$  the energy of a vibrating particle increases and the particle passes over to another energy level (see Fig. 279). Since the potential curve is asymmetrical, the average position of an atom is displaced to the right. Therefore, the average separation between atoms will be greater than the equilibrium separation between atoms at rest, i.e., the  $r$  corresponding to the minimum of the potential well. Thermal expansion results from the fact that the average separation between atoms increases with temperature.

The thermal expansion of a crystal is anisotropic, i.e., in different directions the coefficient of linear expansion  $\alpha$  has different values. Therefore, when indicating the value of  $\alpha$  the crystallographic direction of interest should be specified.

When it is required to give a complete description of the thermal expansion of a crystal, an expansion diagram may be used. Such a diagram is shown for a naphthalene crystal in Fig. 280;  $a$ ,  $b$ ,  $c$  are the crystal axes and  $A_I$ ,  $A_{II}$ ,  $A_{III}$  the axes

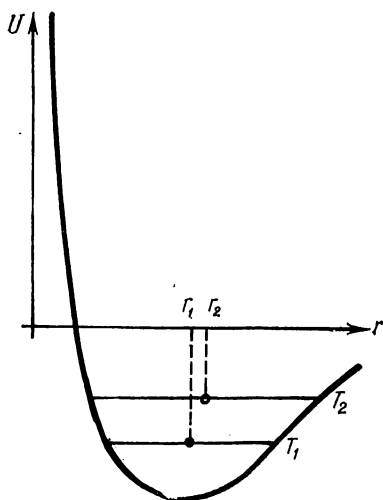


Fig. 279

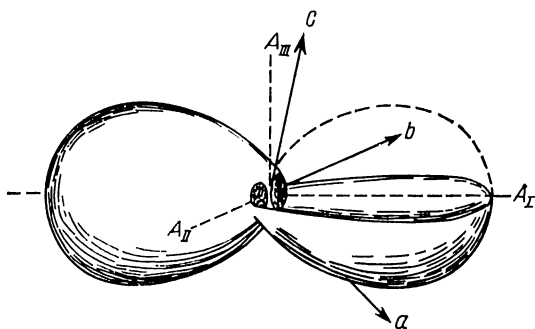


Fig. 280

of symmetry of the expansion diagram. The length of a radius vector drawn from the origin to a point on the surface of the diagram gives the value of  $\alpha$  in the given direction.

The shape of an expansion diagram and its orientation relative to the cell axes accord with the symmetry of the crystal. It cannot be otherwise since physical properties must be the same in directions connected by symmetry operations.

In order to determine linear expansion coefficients, we must have at our disposal means to measure very small displacements with a high degree of accuracy. Instruments for measuring thermal expansion are called *dilatometers* (Greek for "expansion meters"). Interference methods (see Secs. 133 and 135) can provide the required sensitivity (several hundredths of a micron or better), but these require perfectly ground specimens, which are very difficult to prepare. Interference dilatometers are very sensitive to vibrations.

In practice, quartz differential dilatometers are used. In such instruments, the specimen to be measured is placed in a cylindrical holder made of quartz glass. At the bottom of the holder a base prism is placed. One end of the specimen rests on this prism. A quartz rod, which transmits the expansion of the specimen to the measuring portion of the instrument, rests on the upper end of the specimen. The displacement of the end of the rod is measured by means of a microscope or rotary mirror. Since the holder as well as the specimen expands upon heating, the instrument

Coefficients of Linear Expansion

	$t^{\circ}\text{C}$	$\alpha \times 10^4, \text{K}^{-1}$
Aluminium . . . . .	0-100	0.238
Gypsum . . . . .	12-25	0.025
Quartz, $\parallel$ to the axis . . . . .	40	0.0781
Quartz, $\perp$ to the axis . . . . .	40	0.1419
Ice . . . . .	-10-0	0.507

registers the difference between the coefficients of expansion of the specimen and the holder. The necessary corrections, based on available data on the thermal expansion of quartz over a wide range of temperatures, may then be introduced.

The most accurate method of obtaining an expansion diagram is by means of X-ray structural analysis—measurement of the displacements of diffraction spots.

## Sec. 247. CRYSTAL IMPERFECTIONS

**Block Structure.** The structure of an actual crystal differs considerably from that of an ideal space lattice. This assertion is based on numerous facts, including direct electron-microscopic observations. However, our basic knowledge of inner crystal imperfections is derived, in the first place, from strength measurements. A crystal will rupture when it is subjected to a stress of less than a hundredth of the stress which an ideal object should withstand. Deformations and the rupture of crystals will be discussed in Chapter 34. Here, we shall summarise our present knowledge of crystal imperfections.

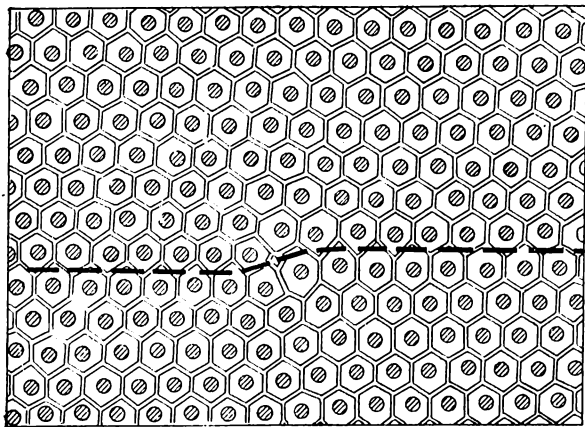


Fig. 281

ary between two blocks. There is good reason to assume that a liquid surface covered with soap bubbles can serve as an excellent model.

An examination of Fig. 281 shows that a "fracture" exists in the atomic rows close to the centre of the model. The portion of the "structure" illustrated in the figure can be represented as four blocks with a common angle at the centre of the model. An imperfection is clearly visible at the centre. Here, the upper "atoms" have not fallen into their right places, i.e., they do not fit into the empty spaces of the compact packing structure. This fault has led to the splitting up of the crystal into blocks.

Many such faults—called dislocations—exist in a crystal. They are distributed randomly and may turn an atomic row to the left or to the right. Therefore, on the average, all crystallographic alignments extend through an entire monocrystal with great exactitude. Dislocations give a monocrystal a block (or mosaic) structure. To be sure, the presence of chance microfissures, or empty spaces, several atoms deep, also facilitate the formation of block structures.

An examination of the figure shows that a dislocation may also be pictured as follows: two adjacent rows, one of which has one particle more than the other—an extra atom has "found its way" into one of the rows.

Briefly, the situation can be described as follows: a monocrystal does not constitute a single lattice. It consists of a large number of tiny blocks which are slightly misaligned (within the limits of several seconds or minutes) with respect to one another. The dimensions of the blocks may vary within rather broad limits. In most cases, they lie in the range of  $10^{-6}$ – $10^{-4}$  cm. Plotting of the block dimensions of a crystal would probably yield a characteristic distribution curve.

Of great interest is the arrangement of particles at the bound-

**Dislocations.** Fig. 281 represents a two-dimensional model of a crystal. It is as if each row were the projection of an atomic layer which is oriented perpendicular to the figure. The large fault would correspond in a three-dimensional crystal to a linear region perpendicular to the figure. This region may be called *the core of the dislocation*. The term indicates that imperfection was caused by "displacement" of one part of the crystal relative to the other.

Dislocation patterns not only explain the block structure of crystals, but also many other phenomena. Therefore, it is well to study these peculiar distortions of crystals in detail.

There are two kinds of dislocations—simple and spiral. The dislocation illustrated by the bubble model is of the simple kind. Schematically, such a dislocation is illustrated in Fig. 282a. The core of the dislocation is designated by an inverted *T*. The distortion is maximal near the dislocation plane dividing the crystal into two parts and rapidly diminishes in either direction away from the dislocation line. Fig. 282b shows a top view of the two adjacent atomic planes on either side of the boundary between the blocks. The upper, or compressed, plane (designated by solid lines) contains one row more than the lower one (designated by dotted lines).

Analogous diagrams for a so-called spiral dislocation is shown in Fig. 283. The lattice is divided into two blocks, one of which has slipped, so to speak, a distance of one period relative to the other (see Fig. 283a). At the axis shown in the figure,

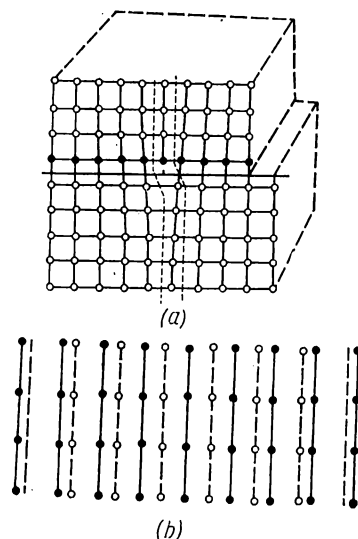


Fig. 282

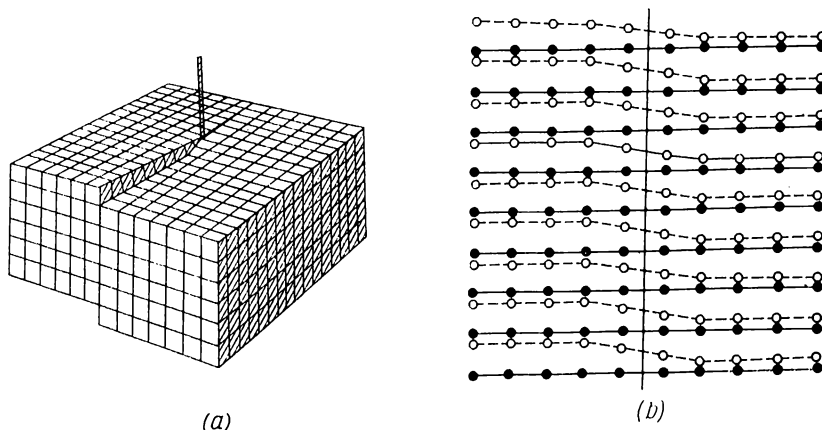
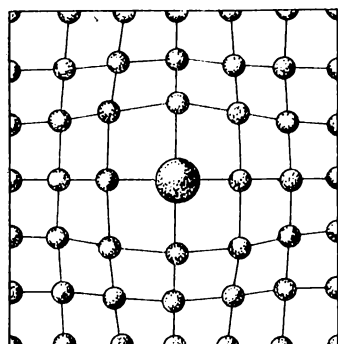


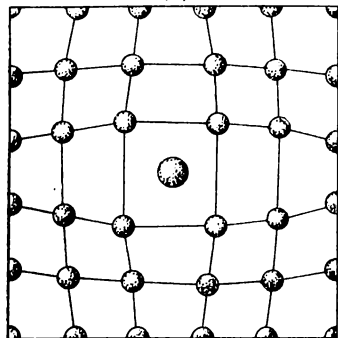
Fig. 283

the distortion is maximal. The region adjacent to this axis is called a region of spiral dislocation. It will be easier to grasp the essence of this distortion if we examine the diagram of the adjacent atomic planes on either side of the boundary between blocks (see Fig. 283b). This is the right view of the three-dimensional

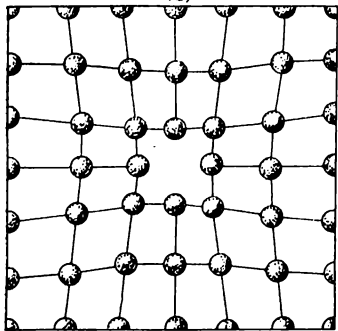
figure. The spiral dislocation axis is the same as in the three-dimensional figure. Solid lines denote the plane of the right block and dotted lines those of the left. As



(a)



(b)



(c)

Fig. 284

may be seen from the diagram, a spiral dislocation differs from a simple dislocation. There is no extra row of atoms in this case. The distortion consists in the fact that the atomic rows change their closest neighbours near the dislocation axis, i.e., they bend and drop down one level.

Why is this called a spiral dislocation? This can be explained as follows. Let us move around the dislocation axis—along the nodal planes of the lattice—beginning at the lowest plane. After each revolution, we are one level higher. In this manner, we reach the top of the crystal, much the same as having climbed a spiral staircase. In Fig. 283, the spiral motion would be counterclockwise. If the blocks were displaced in the opposite direction, the spiral motion would be clockwise.

In a given object, one may encounter successive spiral dislocations having the same rotation direction. If two dislocations having different rotation directions are in the same plane, the resulting distortion is more complex.

**Imperfections within a Block.** A crystal lattice may consist of blocks which also have imperfections. These imperfections may be in the form of lattice vacancies or foreign atoms. A very small number of lattice vacancies and foreign atoms can result in considerable distortion.

Fig. 284 shows the nature of these distortions. In (a) we see the effect of a foreign atom that replaces one of the basic atoms of a lattice, in (b) the effect of a foreign atom penetrating between basic atoms, and in (c) the effect of a lattice "vacancy". The disturbance may extend to a depth of 5-10 lattice spacings in each direction. This encompasses a region of the order of 1,000 cells. Therefore, an impurity of the order of 0.1 per cent may fundamentally change the properties of a crystalline substance. (It should be noted, however, that an impurity does not produce appreciable lattice distortion.) In Sec. 285 we shall

deal with the case of semiconductors in which impurities of the order of one part in a thousand million may change the electrical properties of a body.

#### Sec. 248. SHORT-RANGE ORDER. LIQUIDS

It was seen at the beginning of this chapter that very many solids can be represented as compact packings of spheres. In such structures, the fraction of the total volume consisting of empty space is equal to 26 per cent.



Copper is an example of such a crystal. How does the structure of a piece of copper change when it is melted? Experiments show that the volume increases by about 3 per cent. This increase is due to an increase in empty space, which now equals 29 per cent of the total volume instead of 26 per cent. The compact structure has loosened and the spheres are able to move away from their "proper" positions. The ideal order which is characteristic of a crystal has been disturbed.

As a result of thermal motion, the spheres vibrate, in general, about their equilibrium positions and remain surrounded by the same neighbours. Now and then, neighbours may change when a space of the same size as the volume of a sphere happens to be created in the vicinity of the sphere. Owing to the closeness of particles in a liquid, a so-called short-range order arises. In a model of spheres, one sphere cannot approach another closer than the diameter of a sphere. Such a deviation from ideal randomness occurs in gases as well, but it is of little importance there since the closest molecules in a gas are separated, on the average, by a distance ten times as great as the dimensions of a molecule.

Let us examine a molecule in a liquid and imagine two concentric spheres about it. Assume the radius of one is equal to that of the molecule and the radius of the other to three times this value. On the average, how many close neighbours does such a molecule have? By close neighbours we mean molecules located in the region between concentric spheres. Consider the example of copper, the volume of which increases by 3 per cent when it is melted. According to calculations there are, on the average, 11.6 atoms in the region under consideration. Thus, there are to be found about 12 close neighbours the centres of which are separated from the given atom by a distance equal to its diameter. Closer neighbours are not to be found.

It is clear that the short-range order affects not only close neighbours but successive ones as well. Therefore, it is customary to describe the short-range order by the average density of radial distribution of atoms.

Let us imagine two concentric spheres of radius  $r$  and  $r + dr$  about an atom. For simplicity, assume we are dealing with a monatomic liquid. The volume of this spherical shell will be  $4\pi r^2 dr$ . The number of atoms falling within this shell may be expressed as

$$U(r) \times 4\pi r^2 dr,$$

where  $U(r)$  is the density of radial distribution of atoms.

A  $U(r)$  curve for amorphous arsenic is shown in Fig. 285. Maxima of the curve indicate that certain interatomic spacings acquire considerably more "weight" than others. The origin of successive maxima is exactly the same as of the first. The density of packing is such as to allow the number of closest neighbours of a given atom to fluctuate only within very narrow limits, while the number of neighbours closest to the closest may fluctuate within somewhat broader limits. As the distance from the central sphere increases the random order becomes more and more evident. The  $U(r)$  curve approaches a limit, i.e., the short-range order fades away and gradually passes over into a random order. It is convenient to set the value of  $U(r)$  equal to unity at  $r \rightarrow \infty$ . This distinctive order with regard to close neighbours, which fades away as the distance from the atom or molecule under consideration increases, is what is meant by a short-range order.

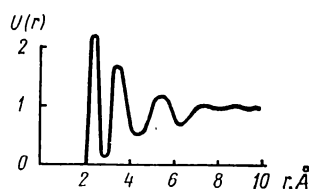


Fig. 285

The order in the arrangement of particles characteristic only of crystals is called a long-range order. This means that the three-dimensional periodicity peculiar to a crystal does not fade away at great distances. The arrangement of atoms along a nodal line regularly repeats itself thousands and millions of times.

When we discussed the structure of crystals, we saw that atoms by no means always behave like spheres. This applies to liquid structures as well.

In an ideal case, the short-range order in an atomic liquid should result in the number of close neighbours being equal to almost twelve. Experiments show that metals the crystal structures of which consist of compact arrangements of spheres continue to have such a short-range order, i.e., an average number of close neighbours just short of twelve, after being melted.

As indicated above, every atom of lithium, sodium and potassium in a crystal has eight close neighbours. The same short-range order is preserved in a liquid, but the average number of close atoms becomes somewhat greater than eight.

Simple substances in the crystalline state of which the atoms are firmly bound to a small number of neighbours behave differently. These bonds are broken when such substances are melted and the number of close neighbours per atom of fluid becomes greater than in a crystal of the same substance.

#### Sec. 249. AMORPHOUS BODIES

The word "amorphous" means "without form". Amorphous solids are the antitheses of regular polyhedral crystals. However, the shape of a polycrystalline body is not regular even though it is not amorphous. How then may crystals and crystalline bodies be recognised? They may be recognised, primarily, by their well-defined melting points. If heat is applied to a crystalline body, the temperature of the body increases until it begins to melt. Thereupon the temperature ceases to rise and the entire melting process takes place at the melting temperature.

Ordinary glass is a typical amorphous solid. It grows soft when it is heated and gradually goes over into the liquid state as the temperature is raised.

This behaviour of amorphous bodies can be explained by their structural peculiarities, leading to the classification of such bodies as liquids rather than solids.

As indicated, there exists a long-range order of particles in crystalline bodies. In amorphous bodies, only a short-range order of particles occurs and in this respect such bodies do not differ from liquids.

Fig. 286*a* shows the structure of quartz (silicon dioxide), and Fig. 286*b* the structure of quartz glass. From the chemical viewpoint, a substance can be obtained in a crystalline form as well as in an amorphous form. The similarities and differences between these two states can be clearly seen in the figure. Apparently, an amorphous body is a "damaged" crystal. In a crystal and an amorphous body, the number of close neighbours and the nature of the encirclement are the same. Possibly, a pentagonal ring is particularly advantageous energetically for  $\text{SiO}_2$  groups. Since the symmetry of an axis of fifth order cannot produce a periodic structure (see p. 468), amorphous glass is obtained.

The absence of a long-range order, which is characteristic of crystalline bodies, is the immediate cause for the absence of a well-defined melting point. At the melting point, a transition occurs and the long-range order disappears. Only a short-range order in the arrangement of atoms remains.

In amorphous bodies, the nature of the arrangement of atoms does not change when the temperature is increased. Only the mobility of the atoms changes, i.e., the atomic vibrations increase. At first only a few atoms are able to escape from

their encirclement and change neighbours. This number gradually increases until, finally, the rate of such changes becomes the same as in water.

The ease with which a given molecule may change its neighbours is related to an important property of a liquid, namely, its viscosity. The less frequently neighbours are changed in a liquid, the thicker, i.e., the more viscous, the liquid. Of

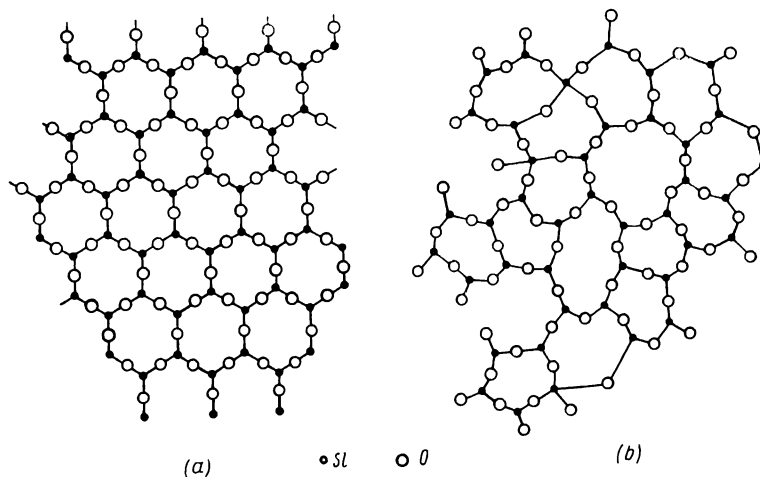


Fig. 286

course, an increase in temperature, which increases the swing of molecular vibrations, results in a decrease in viscosity. It is also quite understandable that, temperature conditions being equal, the liquid whose molecules are more complex will be more viscous. Many liquids harden before they become very viscous. The high viscosity of glue, honey, tar and oil is due to the complex form of their molecules.

When a liquid hardens, the exchange of molecules practically ceases.

#### Sec. 250. SHORT- AND LONG-RANGE ORDER OF ATOMS IN ALLOYS

When two or more substances crystallise together, they may in certain cases form a common crystal lattice. It depends on the relative values of the energy of interaction between homogeneous and heterogeneous particles whether such a mixed crystal is formed or not. If the attraction between homogeneous particles is greater than between heterogeneous particles, a mixed crystal is not formed.

Metal alloys, which are widely used in various industries, are mixed crystals. By reference to the structure of alloys, we can clarify the concepts of short- and long-range order.

In the simple case of binary alloys, we may encounter perfectly ordered structures in which a definite cell can be distinguished and the substance described as a crystal of a compound with the definite formula  $A_nB_m$ . However, this does not always occur and in a number of cases  $A$  atoms randomly replace  $B$  atoms in their lattice or, if they are small, randomly become lodged between  $B$  atoms.

We shall discuss only a substitution alloy, namely, iron-cobalt (see Fig. 287). This alloy has a simple body-centred lattice structure. Each atom—iron as well as cobalt—has eight close neighbours. As regards the arrangement of atom centres, an alloy crystal is always perfectly ordered, i.e., the atom centres form the same body-centred lattice under all conditions. The situation differs with respect to the

distribution of iron and cobalt atoms. Let us consider the lattice points of a crystal to be divided into corners and centres of cubes. For perfect order, all corners are occupied, say, by iron atoms and all centres by cobalt atoms (Fig. 287a). The ideal long-range order of such a crystal may gradually deteriorate if atoms begin to occupy "foreign" sites. But as long as the number of atoms located at their "own" sites differs from the number of atoms located at "foreign" sites (Fig. 287b), the crystal may be said to have a long-range order, even though the order is, in part, "impaired". The long-range order disappears when the distinction between "foreign" and "own" sites is lost, i.e., when half the atoms are at their "own" sites and half at "foreign" sites (Fig. 287c).

It is important to note that when a completely ordered crystal is heated, the order is gradually destroyed, i.e., the percentage of atoms at "foreign" sites increases. There exists a temperature above which even a partially "impaired" long-

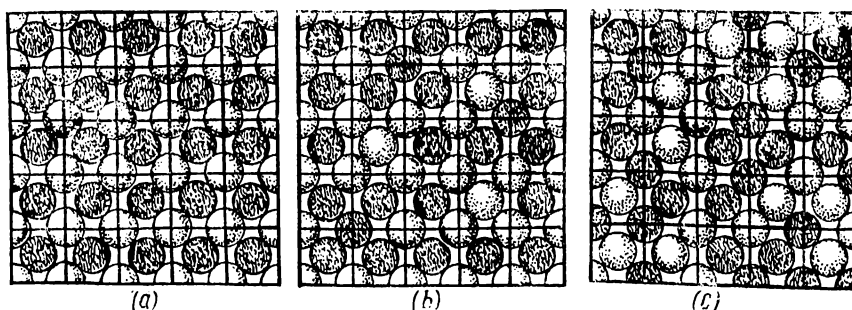


Fig. 287

-range order cannot exist. This temperature is called the  $\lambda$ -point (*lambda-point*). For an iron-cobalt alloy, the  $\lambda$ -point is 770°C. The transition from order to disorder signifies that thermal motion has gained the upper hand over the "tendency" of atoms to maintain a long-range order.

There is a great deal in common between the process of eliminating the distinction between "foreign" and "own" sites and the melting process. Both processes result in the disappearance of a long-range order. However, melting results in the disappearance of the long-range order of atom centres, while passing through the  $\lambda$ -point results in the disappearance only of the order in the arrangement of atoms of different elements.

The basic characteristic of the structure of alloys of the iron-cobalt type is the possible existence of a partial long-range order. Such a partial long-range order can exist only with respect to the distribution of the iron or cobalt atoms, but not with respect to the arrangement of atom centres.

Like in the case of melting, the elimination of a long-range order does not mean the elimination of order in general (a short-range order remains).

The short-range order with respect to the distribution of atoms in iron-cobalt crystals consists in the "tendency" of cobalt atoms to surround themselves with iron atoms, and vice versa. If we take any atom and examine its eight close neighbours, we will find that the number of atoms of the other element will not be equal to one half of the total, i.e., four. Depending on the degree of perfection of the short-range order, an iron atom may be surrounded, on the average, by five, six or seven cobalt atoms.

The investigation of a copper-gold alloy shows that its short-range order has a high degree of perfection. This pertains not only to the number of closest neighbours, but to the number of closest to the closest, etc. If a number of spheres are drawn about any of the gold atoms, it is found that the first shell contains, in effect, only copper atoms, while the second contains only gold atoms. In successive shells, the short-range order begins to deteriorate, but a predilection for atoms of a definite element will be felt as far away as the tenth shell!

It has been determined by means of very precise investigations using X-rays how a long-range order in alloy crystals is "created". Experiments with cobalt-platinum alloys have shown that the regions of long-range order grow in a disordered crystal as crystal nuclei grow in a liquid. These embryonic regions are arranged in a perfectly definite manner relative to the axes of a crystal.

### Sec. 251. LIQUID CRYSTALS

To find examples of liquid crystals, one must turn to organic substances. Molecules of substances forming liquid crystals are always elongated. Liquid crystals are encountered among viruses, and also among lipoids, which are components of living tissue.

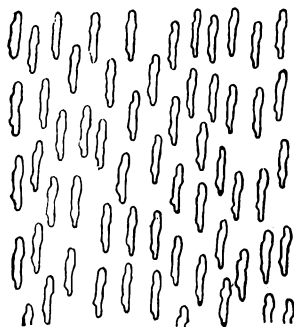


Fig. 288

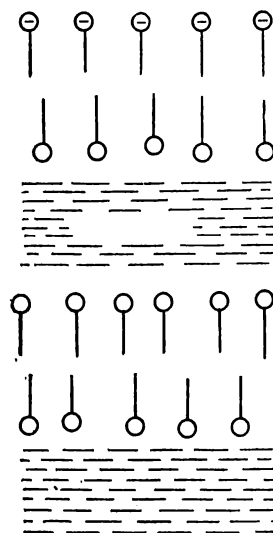
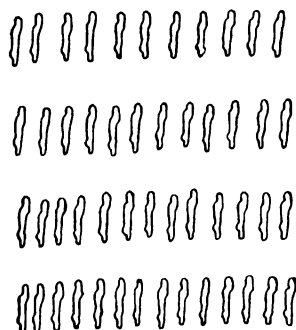


Fig. 289

A substance forming liquid crystals exists as such in a definite temperature range. If a liquid crystal is heated, it turns into an ordinary liquid, if it is cooled, it becomes a crystal.

The term "liquid crystal" is derived from the strange manner in which the properties of a liquid and a crystal are combined. A liquid crystal possesses fluidity and forms drops. However, these drops may be elongated rather than spherical, and somewhat resemble jelly. Careful investigation shows that the order of molecules in such a drop is not like in ordinary liquids.

Two kinds of liquid crystals are known. In one, the molecules are in a short-range order and parallel to one another. In the other, the order of molecules is even more peculiar. Here, the molecules are arranged in layers. Each layer con-

sists of parallel molecules, which are in a short-range order. This is illustrated in Fig. 288.

A soap solution consists of liquid crystals. The washing properties of soap are directly related to its ability to form liquid crystals. A molecule of soap has an elongated shape (transverse axis measuring about 4 Å and longitudinal axis 30-40 Å). At one end of a molecule a negative electric charge is concentrated (and this pole is attracted by water molecules. A soap solution is a liquid crystal. It consists of a large number of double layers of molecules, separated by layers of water (see Fig. 289). In the double layers, the poles of the molecules are turned outward, i.e., toward the water. The molecules of soap within a layer are in a close arrangement and in a short-range order. If there is little soap in the water, the double layers of soap molecules are separated by large layers of water. As more soap is put into the water, more and more double layers will be produced. The solution becomes saturated when the thickness of a water layer equals about 20 Å. The double layers forming a liquid crystal possess great mobility. When we wash our hands, the layers slide easily relative to one another and the skin. Dirt from our hands collects at the poles of the molecules and is then released in the water.

#### Sec. 252. POLYMERS

A large number of organic substances, composed of giant molecules consisting of thousands of atoms, have a peculiar structure. Such substances include plastics, caprone, artificial silk. The molecules of these substances consist of identical groups of atoms arranged in a chain (whence the term "polymer"). The atoms within a molecule are frequently in a long-range order.

The properties of high polymers having lateral chemical bonds between chains differ significantly from those of so-called linear polymers, in which such bonds do not exist. Polymers having lateral chemical bonds between chains are rigid systems. Their atoms are arranged rather loosely and in a completely haphazard manner. Plastics of such polymers are used in the manufacture of buttons, kitchen utensils and various fittings. Linear polymers have interesting properties and structures. Although a number of points still remain unclear, the basic structural features of these substances are well established.

In the solidification of a melt, or directly in the chemical process, long molecular chains become arranged in parallel streams. Since solidification begins simultaneously from a large number of centres, numerous stacks of chains, which collide during growth and pass round one another, are formed. As a result, the final shapes are rather odd and intricate.

The role of a stack in a linear polymer is somewhat analogous to the role of a crystalline particle in a polycrystalline substance. Nevertheless, there is a significant difference between the two, for the degree of order of a group of parallel chains forming a stack (consisting of thousands or tens of thousands of such chains) may vary considerably from case to case. Fig. 290 shows three kinds of order: (a) crystal type—the axes of the chains form a perfect lattice and the azimuths of the chains are ordered, (b) gas-crystal type—the axes of the chains form a lattice and the azimuths are disordered; and (c) amorphous or liquid type—no lattice is formed (absence of long-range order). It should be noted that the displacements of the chains relative to one another in the longitudinal direction also may be ordered or disordered. Since a stack of chains, each of which consists of hundreds of thousands or millions of atoms, is very long and will constitute, therefore, a broken formation, it is evident that even in the case of ideal order in the arrangement

of chains the order of a stack is not entirely like that of a crystal. Only the portion of a stack for which the parallel chains are rectilinear can have a crystal arrangement. This means that in the case of ideal order each stack of chains consists of a sequence of crystalline regions (crystalline particles).

The stretching of linear polymers consists in the unfolding of stacks of chains. A similar mechanism of elongation enables us to explain extensions of up to 1,000 per cent occurring in certain natural and artificial high polymers.

Rubber and polyethylene are the best known linear polymers. The high polymers used in the manufacture of artificial fibre also have the same kind of structure.

#### Sec. 253. BIOLOGICAL MACROMOLECULES

The molecules responsible for vital activity of an organism are also linear molecules consisting of hundreds of thousands and millions of atoms.

The nucleus of a cell contains a molecule of desoxyribose nucleic acid (DNA) which is the bearer of heredity. These molecules exist in the form of a double helix. Fastened in a certain order to the basic chain of atoms which is the backbone of the helix are molecular radicals of four types. Genetic information is coded by the order in which they follow along the chain. Such a chain contains about  $10^6$  radicals. The number of possible permutations of four elements in such a long chain is unimaginably large. It is now clear what a rich amount of information is borne on a long macromolecule, and that not only the peculiarities of a biological species, but also the peculiarities of a certain individual can be coded by the order in which the radicals follow in a DNA molecule.

During cell division double helices get untwined and, passing on to posterity, hand it down the ancestral features by means of the code of sequence of molecular radicals.

Double helices of DNA can be singled out and their structure investigated. But today we know how to determine the sequence of atoms and atom groups only in ordered crystalline structures. The molecule of DNA is, in principle, unordered, and that is why, using physical methods (spectral and, mainly, X-ray technique), investigators have succeeded only in establishing the principles of its construction. The development of methods of establishing the sequence of radicals in DNA for a given organism, i.e., for objective describing of ancestral features on a molecular level, still remains a problem.

Double helices of DNA can be singled out and their structure investigated. But today we know how to determine the sequence of atoms and atom groups only in ordered crystalline structures. The molecule of DNA is, in principle, unordered, and that is why, using physical methods (spectral and, mainly, X-ray technique), investigators have succeeded only in establishing the principles of its construction. The development of methods of establishing the sequence of radicals in DNA for a given organism, i.e., for objective describing of ancestral features on a molecular level, still remains a problem.

The cell—a basic living “brick”—is a “factory” producing protein molecules which fulfil various vital functions. Protein molecules are produced, so to say, under the leadership of the DNA molecule. They are constructed from twenty types of amino acids, the order of their combination strictly defined for the protein of a given type being dictated by the molecule of DNA. The latter plays the role of a

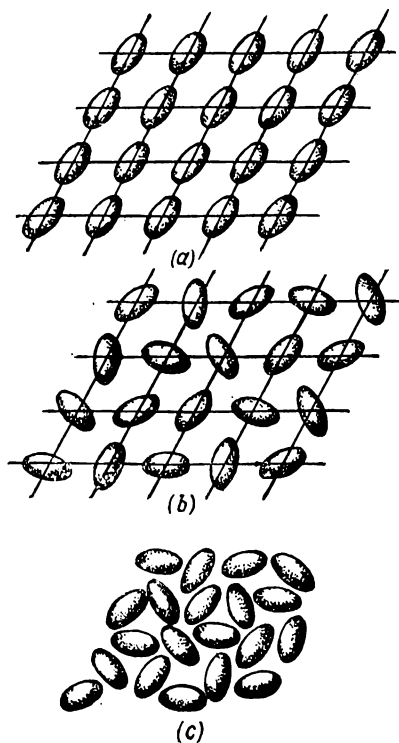


Fig. 290

linear combination of hundreds and thousands of matrices "printing" different protein molecules. Some of them are similar for a given biological species, others are affected by the peculiarities of a certain individuum.

Protein molecules can be singled out and crystallised. Despite the fact that their molecules contain thousands of atoms, protein crystals can be successfully studied using the methods for X-ray structural analysis. At present we are able to determine their structure. This problem is extremely complicated, that is why there is no wonder that today we know the structure of only six types of proteins.

The fact of formation of crystals of such a complex molecule is remarkable by itself. Fantastically branched sequence of several thousands of atoms is completely identical in all billions of the molecules forming a protein crystal. All the molecules are arranged in one-two orientations at the points of a regular three-dimensional lattice.

A protein crystal possesses the following remarkable property: the molecules of which it is made up touch one another only with a small fraction of their surface, the most part of intermolecular space being filled with water. When carefully dried, protein crystals undergo transformations which change the arrangement of their molecules. The most part of water can be removed with crystallinity of the protein retained.



## Phase Transformations

### Sec. 254. PHASE DIAGRAMS

A substance can have only one gaseous state and one liquid state, but it can have several crystalline states (also, several liquid-crystalline and gaseous-crystalline states).

The gaseous and liquid states of a substance are characterised by disorder in the arrangement of particles. In a gas, the ratio of the kinetic energy of the particles to their potential energy of interaction is such that the binding forces cannot restrain the particles from flying apart to the extent to which the vessel containing the gas permits. The liquid state has a definite form since the binding forces do not permit the molecules to have independent free paths.\* At high pressures, the distinction between a gas and a liquid disappears.

Since two basically different random arrangements of particles do not exist, every substance has one liquid and one gaseous state. A crystal is characterised by a definite arrangement of particles and, in principle, an infinite number of different crystal phases can exist for a given substance. In actuality, several different crystal phases exist, as a rule, for one and the same chemical compound (diamond and graphite, white and gray tin, yellow and red sulphur, etc.).

Every substance exists in one or another phase, depending on the external conditions, viz., the temperature and the pressure. It is customary to use a pressure-temperature diagram, instead of a table, to describe the conditions for the existence of the various phases of a given substance. A diagram of this kind is known as a *phase diagram*.

Three such diagrams are shown in Fig. 291. In the upper left-hand corner, we see a phase diagram for an ideal substance having only one solid phase. The diagram is divided into three regions: one region indicates the conditions for the existence of a crystal, another for the existence of a liquid, and the third for the existence of a gas. The gaseous state, of course, is represented by the lower right-hand portion of the diagram, i.e., where the temperatures are high and the pressures low. The solid phase is represented by the region of lowest temperatures and highest pressures. Such a diagram is very convenient. In order to determine the state of a body at a pressure  $p$  and a temperature  $T$ , all we need to do is to find this point on the diagram and observe in which region it is located.

In the upper right-hand corner of Fig. 291, the conditions for the existence of the various phases of sulphur are shown. This substance has two crystal phases and therefore the diagram is divided into four parts. In the lower part of Fig. 291, the phase diagram for water is shown. Since it is difficult to represent such a diagram drawn to linear scale on a single drawing, the pressure in this diagram has been plotted logarithmically. It will be seen that ice exists in five different phases, which are designated by the Roman numerals I, II, III, V and VI. The phase which was originally designated as IV turned out to be an error. The less common phases of ice exist at higher pressures.

On a phase diagram, boundary lines between phases, as well as points within the various regions of the diagram, have physical significance. At pressures and temperatures corresponding to points on the boundary lines, two boundary phases exist simultaneously. In the case of water, this would correspond to the condition

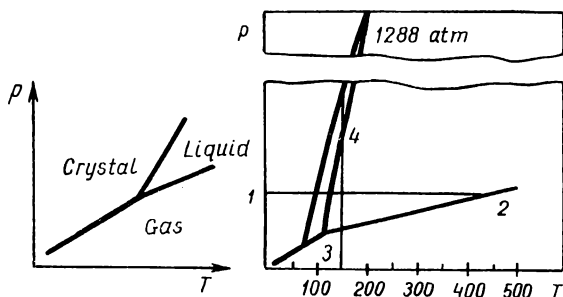
---

\* In the absence of gravitational forces, a drop of liquid is spherical.

of ice floating on water, with the ice not melting and the water not freezing. The boundary lines may be called *phase equilibrium curves*.

It should be emphasised that these are phase equilibrium curves and not points. This means that equilibrium between two phases can be realised at different tem-

peratures if the pressure is varied accordingly. This may also be expressed as follows: the temperature of phase equilibrium is a function of pressure; or, the pressure of phase equilibrium is a function of temperature.



#### Sec. 255. PHASE TRANSFORMATIONS

Phase equilibrium curves may also be called phase transformation curves since transition from one phase to another occurs when this line is crossed.

The boundary line between a solid and a liquid is the *fusion*, or *crystallisation*, curve; and the boundary line between a liquid and a vapour is the *vaporisation*, or *condensation*, curve. We call the boundary line between a solid and a vapour the *sublimation curve* and the lines between two solid phases simply *transformation curves*.

Processes involving a change of state are also conveniently indicated on a phase diagram. Usually, we are con-

cerned with transformations occurring at constant temperature or at constant pressure. These processes are represented on a diagram by vertical and horizontal lines, respectively.

Several phase transformations are illustrated in Fig. 291. Line 2-1 on the phase diagram of sulphur represents a cooling process for a sulphur gas at normal pressure. At a temperature of 444.5°C sulphur is transformed from a gas into a liquid, at 140.2°C from a liquid into a crystal phase and, finally, at 95.5°C from this crystal phase into another crystal phase. The compression of sulphur gas is illustrated in the same diagram by the process 3-4. By increasing the pressure, we are able in this case too to transform a gas into a liquid and then, at very high pressures (above point 4) into solid states.

Under certain unique conditions, three phases may exist together simultaneously. Such points are called *triple points*. Sulphur has three triple points: (1) gaseous, liquid and yellow sulphur existing simultaneously; (2) gaseous, liquid and

Fig. 291 .

red sulphur existing simultaneously; and (3) a liquid existing in equilibrium with the two crystal phases.

It can be proved on the basis of strictly thermodynamic principles that a quadruple point cannot occur. Thus, no conditions exist under which, for example, two crystal phases are in equilibrium with a liquid and a vapour.

Every phase transformation is characterised by a transition temperature at a given pressure. We speak of the melting (crystallisation), boiling, sublimation, etc., points of a substance. If the pressure is not indicated, this usually means that the transformation occurs at normal atmospheric pressure.

An important characteristic of a transformation is its heat of transition. The occurrence of latent heat of vaporisation and heat of fusion is well known, but heat of transition is a general phenomenon. A transformation which proceeds by heating absorbs heat. In accordance with the second law of thermodynamics, heat of transformation is uniquely related to a change in entropy:

$$\Delta Q = T \Delta S,$$

where  $T$  is the transition temperature. Therefore, it is evident that a phase transformation which proceeds by heating is accompanied by an increase in entropy.

The transition (fusion, boiling) temperature can be calculated from the formula

$$T = \frac{\Delta Q}{\Delta S}$$

i.e., it is equal to the latent heat of transition divided by the increase in entropy. But in this form the statement is of purely theoretical significance since practically the change in entropy for a phase transformation cannot be pre-calculated. However, knowing the transition temperature and heat of transition from experiments, one can determine accurately the magnitude of the increase in entropy.

The phase transition from ice I to ice III occurs at a temperature  $t = -20^\circ\text{C}$  and a pressure  $p = 2,403$  atmospheres. This transition is accompanied by the release of heat; each gram of ice releases  $\Delta Q = 5.6$  cal. Therefore, the change in entropy is

$$\Delta S = \frac{\Delta Q}{T} = \frac{5.6}{253} = 0.022 \text{ cal/K.}$$

#### Sec. 256. THE PHASE DIAGRAM AND PROPERTIES OF HELIUM

Helium is worthy of a separate description, since only this element reveals two exceptions from the general rules. Its phase diagram (isotope 4) is shown in Fig. 292. It turns out that there exist two solid phases. The body-centred lattice of helium is stable in a very narrow temperature-pressure region. In the principal region, helium has a close-packed hexagonal structure (see Sec. 242), and at very high pressures (not shown in the diagram) it acquires a face-centred (cubic) lattice structure.

The first striking peculiarity is that at zero pressures and a zero temperature liquid but not solid is a stable phase. The second peculiarity consists in that liquid helium can exist in two states separated by a distinct phase boundary.

These peculiarities of helium are due to a combination of a small mass of atoms with extremely weak forces of interaction. It is sufficient to indicate that the depth of the potential well on the curve of interaction of two helium atoms is one tenth of that for argon. As a result, it turns out that zero energy of helium, i.e., the kinetic energy in the lowest state is so high that, without applying an external pressure, a helium atom cannot be found (in contrast to other atoms) in

the potential well created by interaction with the neighbouring atoms and confine itself in its motion to the vibrations about the equilibrium state.

The most striking property of helium II is superfluidity, i.e., a complete absence of viscosity, which was discovered by the well-known Soviet physicist P. L. Kapitza in 1938.

Superfluidity can be observed in a specially designed vessel whose bottom is fitted with a very narrow slit—only half a micron wide. An ordinary liquid almost does not infiltrate through such a slit. At temperatures higher than 2.19 K helium behaves in a similar manner. But as soon as the temperature becomes lower than 2.19 K, the rate of discharge of helium increases abruptly at least several thousand

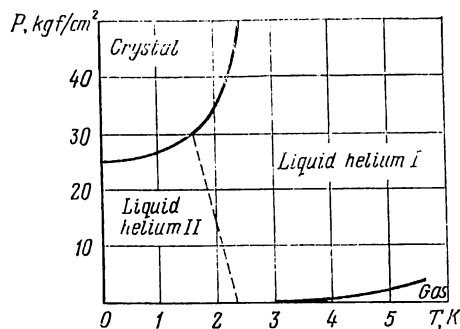


Fig. 292

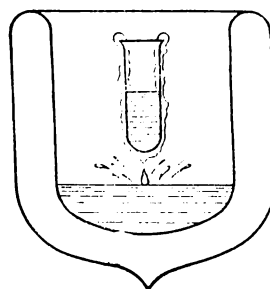


Fig. 293

times. Helium II flows out through the narrow gap almost instantaneously, i.e., it completely loses its viscosity. Superfluidity of helium leads to an even stranger phenomenon: helium II is capable of “coming out” (without any external help) of a glass or a test tube into which it was poured.

Figure 293 shows the scheme of carrying out this experiment. A test tube filled with helium II is placed into a Dewar vessel above the helium bath. The helium rises along the wall of the test tube in the form of an extremely thin, imperceptible film, overflows the edge of the tube, and, finally, drips from its bottom.

Due to capillary forces, molecules of any liquid wetting the vessel wall rise along this wall and form on it an extremely thin film whose thickness is equal to one millionth of a centimetre. In the case of an ordinary viscous liquid, this film is imperceptible for man's eye, and manifests itself in no way.

Quite another thing happens when we deal with helium which is deprived of viscosity. A narrow slit does not hinder the motion of superfluid helium; and a thin surface film is just the same as a narrow slit. A liquid deprived of viscosity flows in the form of an extremely thin layer. Over the edge of a glass or a test tube, the surface film forms a siphon through which helium flows off.

Nothing of the kind is observed in the case of ordinary liquid. It is practically impossible for a liquid of normal viscosity to “make” its way through a siphon of insignificant thickness. Such a “motion” is so slow that the overflow would take millions of years.

Thus, helium II has no viscosity whatsoever. Logically, we might draw a conclusion that in such a liquid a solid body must move without friction. Let us place a disk tied to a thread into liquid helium and wind the thread. Leaving alone this simple device, we shall create a kind of pendulum: the thread with the disk will vibrate and periodically wind now clockwise, now anticlockwise. If there is no friction, then we should expect that the disk will vibrate forever. But in a compar-

atively short period of time, approximately the same as for an ordinary normal helium I (i.e., for helium at a temperature higher than 2.49 K), the disk ceases to move. Outflowing through a slit, helium behaves like a liquid deprived of viscosity, but with respect to bodies moving in it, it behaves as an ordinary viscous liquid.

The behaviour of liquid helium can be understood only from the viewpoint of quantum mechanics. We shall try to show how the theory developed by L. D. Landau explains the behaviour of liquid helium.

It turns out that each particle of liquid helium takes part simultaneously in two motions: one is superfluid, the other ordinary.

Helium II behaves in such a manner as if it consists of a mixture of two liquids moving absolutely independently "one through the other". One liquid behaves in a normal way, i.e., it possesses ordinary viscosity, while the second is superfluid.

When helium drips through a slit or flows off over the edge of a glass — we observe the effect of superfluidity. And in the case of a submerged disk, the friction stopping the disk occurs due to the fact that in the normal portion of helium it is inevitable.

The ability of participation in two different motions gives rise to entirely unusual heat conducting properties of helium. As it was already stated, liquids, in general, are poor heat conductors. Helium I is not an exception in this respect. But when transformed into helium II, its heat conductivity increases approximately billionfold. As a matter of fact, helium II conducts heat better than the best ordinary heat conductors, such as copper and silver.

The point is that the superfluid motion of helium does not participate in heat conductivity. Therefore, when there is a temperature drop in helium II, there arise two flows moving in opposite directions, and one of them—the normal one—carries heat. This differs from usual heat conductivity. In an ordinary liquid, heat is transferred by molecule knocks. In helium II heat flows together with the ordinary portion of helium, it flows like a liquid. Such method of heat transfer leads to a tremendously high heat conductivity.

The correctness of the above considerations can be proved directly by conducting the following experiment which is very simple by its idea.

Placed in a bath with liquid helium is a Dewar vessel also filled with helium. The vessel is in communication with the bath through a capillary tube. The helium inside the Dewar vessel is heated by means of an electric coil, but heat is not transferred to the helium contained in the bath since the walls of the Dewar vessel do not conduct heat.

A wing suspended from a thin thread is fitted opposite the capillary tube. If heat flows like a liquid, then it must turn the wing. And this is just what happens; the amount of helium in the vessel remaining unchanged. How can we explain this phenomenon? When the helium is heated, there occurs a flow of the normal portion of the liquid from the heated to the cold place, and an opposite flow of the superfluid portion. The amount of helium at each point remains constant; but since the heat transfer is accompanied by the normal portion of the liquid, the wing is turned due to viscous friction of this portion and is kept deflected as long as the heating lasts.

From the fact that the superfluid motion does not transfer heat, we may draw another conclusion. As we mentioned above, helium "creeps" over the glass edges. But it is only the superfluid portion that "comes" out of the glass, the normal portion remaining in the glass. Heat is associated only with the normal fraction of helium, it never accompanies the "coming out" superfluid portion. Hence, as the helium keeps "coming out" of the vessel, one and the same amount of heat warms

a decreasing amount of helium. As a result, the helium remaining in the vessel must get heated which is actually observed in the course of the experiment.

The masses of helium associated with the superfluid and normal motions are not equal to each other. Their ratio depends on temperature: the lower the temperature, the greater is the superfluid portion of the helium mass. At the absolute zero (i.e., at a temperature of  $-273^{\circ}\text{C}$ ) the whole amount of helium becomes superfluid. As temperature grows, more and more helium behaves in a normal way, and at a temperature of 2.19 K the whole amount of helium becomes normal, acquiring the properties of an ordinary liquid.

#### Sec. 257. PHASE STABILITY

How do we explain the fact that under certain conditions a body constitutes a liquid and under others a solid? There are two tendencies which determine the nature of a state under given external conditions: first, the tendency of a body to possess a minimum of energy and, secondly, the tendency to possess a maximum of entropy. The first tendency is a consequence of the fact that a system of molecules behaves like any system of mass points subject to the laws of Newtonian mechanics and, as we know, a mechanical system tends to possess a minimum of potential energy. The second tendency is a consequence of the second law of thermodynamics.

During the transition from a gas to a liquid and during the transition to a solid, the internal energy of a substance decreases. The energy of a gas is higher than the energy of a liquid since in passing from a liquid to a gas work must be expended in overcoming the binding forces between molecules. And the energy of a crystal is lower than the energy of a liquid since an ordered arrangement of interacting particles is always more stable than a disordered arrangement. This can be proved rigorously, but the proof will not be presented. The statement appears rather obvious. Imagine, for example, a perfect lattice of spheres connected by means of springs. Any displacement of a sphere requires a certain amount of work. Hence, an ordered arrangement corresponds to a minimum of energy.

Entropy behaves in a different manner. Roughly speaking, the greater the freedom of motion of the constituent particles of a body, the greater its entropy. A disturbance in the order or an increase in the separation between particles results in an increase in entropy.

Thus, for a given temperature and pressure, the state of a substance is established as a compromise between entropy and energy. Using the second law of thermodynamics, we can obtain a quantitative expression for this general law.

Imagine that a body is placed under "foreign" conditions, i.e., ice under the conditions for the existence of water, etc. In such a case, an irreversible phase transformation (fusion, vaporisation, etc.) will occur in accordance with the second law of thermodynamics: the increase in the entropy of a body will be greater than the applied reduced heat,

$$dS > \frac{dQ}{T}.$$

Using the first law of thermodynamics, we can rewrite the inequality in the form

$$dS > \frac{dU + p dv}{T} \quad \text{and} \quad dU - T dS + p dv < 0.$$

Since the phase transformation occurs at constant temperature, we obtain

$$d(U - TS) + p dv < 0.$$

If the process is not accompanied by a change in volume, the transition to a state of equilibrium takes place with  $d(U - TS) < 0$ , i.e., with a decrease in the quantity  $F = U - TS$ . This function is called *the free energy*. We have shown that a spontaneous phase transformation is accompanied by a decrease in free energy, i.e., the free energy of a stable state must be a minimum.

If the process occurs at constant pressure, the transition to an equilibrium phase takes place with  $d(U - TS + pv) < 0$ , i.e., with a decrease in the quantity  $\Phi = U - TS + pv$ . This function is called *the thermodynamic potential*. Thus, a phase transformation at constant pressure is accompanied by a decrease in thermodynamic potential, i.e., the thermodynamic potential will have a minimum value at equilibrium.

The opposing tendencies of entropy and internal energy are brought out in this statement: a decrease in energy and an increase in entropy result in a decrease in free energy or thermodynamic potential. These two tendencies have been expressed quantitatively in the relations showing that  $F$  and  $\Phi$  tend to a minimum.

The formulated condition for phase equilibrium has numerous applications. For example, using this condition, we can derive a relation for the slope of a phase equilibrium curve.

On such a curve, consider two points representing the external conditions  $T_1, p_1$  and  $T_2, p_2$ . The equilibrium conditions for these points have the form

$$\Phi_1(T_1, p_1) = \Phi_2(T_1, p_1) \quad \text{and} \quad \Phi_1(T_2, p_2) = \Phi_2(T_2, p_2).$$

The subscripts of  $\Phi$  refer to the phases in equilibrium. Subtracting the first equation from the second, we obtain

$$\Phi_1(T_2, p_2) - \Phi_1(T_1, p_1) = \Phi_2(T_2, p_2) - \Phi_2(T_1, p_1).$$

Let us assume that the two points are close to each other. Then, by means of the formula for the increment of a function of two variables, we can transform the above equation into the form

$$\frac{\partial \Phi_1}{\partial T} dT + \frac{\partial \Phi_1}{\partial p} dp = \frac{\partial \Phi_2}{\partial T} dT + \frac{\partial \Phi_2}{\partial p} dp.$$

Substituting the values of the derivatives of the function  $\Phi = U - TS + pv$ , namely,  $\frac{\partial \Phi}{\partial T} = -S$  and  $\frac{\partial \Phi}{\partial p} = v$ , we obtain

$$\frac{dp}{dT} = \frac{S_1 - S_2}{v_1 - v_2} = \frac{\Delta S}{v_1 - v_2}.$$

But since  $\Delta S = \frac{\Delta Q}{T}$

$$\frac{dp}{dT} = \frac{\Delta Q}{T(v_1 - v_2)} \quad (\text{Clapeyron-Clausius equation}).$$

Thus, the slope of the curve (the derivative  $\frac{dp}{dT}$ ) is determined by the latent heat of fusion  $\Delta Q$ , the temperature of the phase transition  $T$ , and the difference in the volume of the phases. If  $\Delta Q$  is positive, this means that the subscript 1 refers to the high-temperature phase.

Let us apply the Clapeyron-Clausius equation to the case of melting ice. When ice melts, 1 cm<sup>3</sup> of water is obtained from 1.091 cm<sup>3</sup> of ice. The volume change  $v_1 - v_2$  is equal to -0.091 cm<sup>3</sup> (the volume decreases). In this case,  $\Delta Q$  will be the heat of melting and is equal to 80 cal/g. The temperature  $T$  equals 273 K. Hence,

$$\frac{dT}{dp} = \frac{T \Delta v}{\Delta Q} = \frac{(273) \times (-0.091)}{80} = -0.31 \frac{\text{deg cm}^3}{\text{cal}}.$$

Since the dimensions of the above result somewhat obscure its significance, let us convert calories to atmospheres, recalling that  $1 \text{ cal} = 42.7 \text{ kgf cm} \approx 42.7 \text{ atm cm}^3$ . We obtain

$$\frac{dT}{dp} = -0.0075 \frac{\text{deg}}{\text{atm}}.$$

Thus, increasing the pressure by 1 atm decreases the melting point of ice by 0.0075 degree.

## Sec. 258. METASTABLE STATES

The above thermodynamic explanation of phase transition phenomena does not explain a number of observed facts. Thus, from the thermodynamic viewpoint, for a given  $p$  and  $T$  there can occur a single state (a point in one of the regions of a phase diagram) for which the free energy, or thermodynamic potential, assumes a minimum value. However, it is possible for graphite and diamond to exist side by side, and water can be obtained under the conditions for the existence of ice (supercooled water). Numerous other examples in which the above thermodynamic principles are violated may be cited. The situation can be described as follows: in addition to states which are stable under given external conditions, so-called metastable states may also exist.

The free energy of a metastable state is not a minimum, but nevertheless the transition from this state to a state having a minimum energy is impeded. Different metastable states may differ considerably in their degree of stability. Sometimes a slight impulse suffices for a transition to occur to a "normal" state, while in other cases a metastable state may be, in actuality, no less stable than "normal" state.

Various phase transformations can be delayed. Thus, water can be supercooled, i.e., at normal pressure, water may exist at a temperature below  $0^\circ\text{C}$ ; water can also be superheated, i.e., its temperature may be raised above  $100^\circ\text{C}$  without boiling. A vapour also may be obtained under atypical conditions (a supercooled vapour is said to be supersaturated). Transformation delays always occur in the solid state, i.e., the transformation of one crystal phase into a second is delayed even though the conditions prevailing are those for the stable existence of the second phase.

However, one type of transformation, namely, fusion, is never delayed. Thus, a crystal cannot exist under conditions that are stable for the liquid phase.

We frequently have occasion to deal with supercooled liquids. Liquids such as glycerine considerably increase in viscosity when supercooled and may remain in the amorphous state for months or even years. Glass is another example of a supercooled liquid.

The existence of a metastable state can be demonstrated in the case of a supercooled liquid by bringing the liquid into contact with a crystal. In such cases, crystallisation begins immediately. If the liquid is highly supercooled, the effect will be extremely violent. When a snowflake is thrown into supercooled water, ice needles dart through the water in all directions and in a few seconds the transformation is complete.

Delays of crystal-crystal transformations are particularly interesting. Here, delays can occur, so to speak, in both directions. Yellow sulphur should be transformed into red sulphur at  $95.5^\circ\text{C}$ . If sulphur is rapidly heated, this transformation point may be "skipped" and the sulphur may be brought to fusion at a temperature of  $113^\circ\text{C}$ . Now assume that the melt is gradually cooled. At  $113^\circ\text{C}$  small crystals of red sulphur are formed. Cooling does not result in a transformation at  $95.5^\circ\text{C}$ , and even at room temperature small crystals may exist for a considerable



period of time. However, the transformation process proceeds, even though slowly, and within a day is complete, i.e., a yellow powder is obtained. Here, too, the metastable nature of the state is best demonstrated by dropping a small crystal into the melt.

In certain cases, we are interested in a substance in a phase which might be expected to exist under entirely different conditions. An example of this is white tin which is transformed into gray tin when the temperature is reduced to 13°C. Usually, we are more interested in white tin and are cognisant of the fact that in winter nothing can be done with it. However, white tin excellently withstands 20-30° of supercooling and only under severe winter conditions does it begin to be transformed into gray tin. (The members of the Scott expedition to the South Pole perished as a result of ignorance of this fact. Liquid fuel taken on the expedition had been placed in containers soldered with tin. At the extremely low temperatures prevailing, the white tin was transformed into a gray powder. As a result, the containers opened and the fuel was lost.)

To explain transformation delays, let us consider the difference between liquid-crystal and crystal-crystal transformations on the one hand and crystal-liquid transformations on the other. In the last case the long-range order of atoms disappears, while in the first two a long-range order is created. The elimination of long-range order does not require a great effort. Fusion begins at the surface; atom after atom is torn away from its neighbours and falls out of strict order.

On crystallisation, short-range order is transformed into long-range order. The process begins at the surface and must proceed inwardly, i.e., into the substance. The atoms, or molecules, are "forced" to establish strict order under extremely crowded conditions. Their motions must be harmonised for order to be established. As we have seen, the rearrangement of atomic order, which requires that atoms undergo "organised" displacements from certain ordered positions to others, is all the more difficult.

Transformations in the solid state always begin at the boundaries of grains, blocks, empty spaces, and at dislocations; in other words, wherever there is more freedom. If only several score atoms have occupied positions corresponding to a new order, oriented growth of the nucleus proceeds, i.e., one after another atoms begin to pass from the old, less favourable order, or in the case of crystallisation from disorder, to the new order. This is the effect of a crystal particle, or seed, which invariably puts an end to a supercooled state.

#### Sec. 259. GAS ⇌ LIQUID TRANSFORMATIONS

Vaporisation consists in the separation of fast-moving particles from the surface of a liquid. Two conclusions immediately follow from this, namely, vaporisation increases with increasing temperature and requires the application of heat. If the vaporised molecules are continuously removed from the surface of the liquid, the vaporisation process continues until all of the liquid has been transformed into vapour.

Let us consider vaporisation in a closed vessel. In such a case, not only do molecules separate from the surface of a liquid, but the reverse process also occurs, namely, vapour molecules return to the liquid. The vaporisation process will continue until dynamic equilibrium corresponding to the given temperature has been established. Of course, the liquid may completely vaporise without equilibrium being established with the vapour.

When equilibrium exists, we say that a vapour is saturated. The pressure of a saturated vapour is a function of the temperature and is given by the phase

equilibrium curve. By changing the temperature, we either vaporise more of the liquid in a vessel or condense some of the vapour. This results in a change of the vapour pressure.

It is clear why the density and pressure of a saturated vapour increase with temperature. The number of molecules leaving a liquid rapidly increases as the kinetic energy of the molecules increases. On the other hand, the number of vapour molecules returning to a liquid is almost independent of the temperature since such a process requires no energy.

The density of saturated vapour at a given temperature varies from substance to substance within a broad range. At room temperature, the density of saturated steam is equal to 13 mm, while that of saturated mercury vapour is only 0.005 mm.

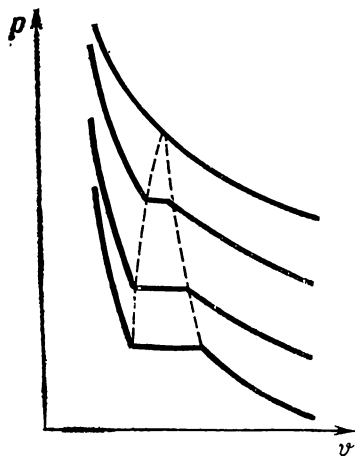


Fig. 294

A clear picture of transition processes from gases to liquids may be obtained by considering isothermal compression of a gas, i.e., "vertical" processes in a phase diagram. In order to represent volumetric changes, which are not shown in a phase diagram, let us draw an auxiliary diagram in which pressure is plotted as a function of volume (see Fig. 294).

If gas compression occurs at a sufficiently low temperature, sooner or later we arrive at an intersection point with a phase equilibrium curve. At this instant, the pressure is equal to that of saturated vapour at the temperature of the experiment and the first drops of liquid appear. As long as a vapour is not completely transformed into liquid, the compressing motion of the piston will not be accompanied by a change in pressure since we remain at the same point in the phase diagram

from the beginning to the end of condensation. The condensation process will be indicated by a horizontal line on the pressure-volume curve.

The significance of points of a rectilinear segment on the diagram is clear. They describe a two-phase liquid-vapour system. Each point of such a segment corresponds to a definite ratio between the phases, which can be easily determined by means of a "lever" rule. Let us designate the volume of the liquid by  $v_1$ , the volume of the vapour by  $v_2$ , and the proportion of substance in the liquid state by  $x$ . Then, the volume of the wet mixture is given by

$$v = xv_1 + (1 - x)v_2.$$

Hence,

$$x = \frac{v_2 - v}{v_2 - v_1}.$$

This is the "lever" rule.

When the condensation process is completed, a steep rise occurs in the curve since liquids have very low compressibility.

Now, let us increase the temperature to that of the next isotherm on the diagram. It will be practically the same as the first, except for one important difference, namely, condensation begins later since the pressure of a saturated vapour is greater at a higher temperature. Moreover, condensation is completed earlier since the piston does not reach its preceding position owing to the thermal expansion of the liquid.

By increasing the temperature further, we obtain a series of isotherms in which the horizontal segments corresponding to two-phase systems become shorter and shorter. Finally, this segment disappears entirely. The decrease in the length of the horizontal segment indicates that the specific volume of the liquid is approaching that of the vapour. At a certain critical temperature, these volumes become equal and the isotherm no longer has a horizontal portion. In the pressure-volume diagram, the critical point is easily determined as the apex of the dotted two-phase region. In a diagram of state, the critical point is located where the liquid-vapour phase equilibrium curve breaks off.

As the temperature is increased, the isotherms resemble broken curves less and less and gradually approach the hyperbolas of an ideal gas.

The existence of a critical point indicates that we were justified in stating that there is no basic difference between a gas and a liquid. We see that if we by-pass the critical point a transition from a liquid state to a gaseous state can be achieved without going through a phase transformation.

#### Sec. 260. LIQUEFACTION OF GASES

We cannot obtain a liquid by compressing a substance the temperature of which is above the critical point. When such a substance is highly compressed, it becomes very dense, with its molecules coming into close contact with one another. Nevertheless, a liquid in the usual sense of the term cannot be obtained. The substance cannot be poured into a glass like an ordinary liquid, i.e., its state has no distinctive form. This is due to the fact that the phase equilibrium curve was not crossed during compression. The absence of such a crossing indicates that a two-phase liquid-gas system cannot be obtained. This means that we cannot get a liquid with a definite form; instead, the liquid fills the entire volume available to it.

Compression must take place at temperatures lying below the critical point if a gas is to be liquefied. This is easily achieved if the required temperatures are reached by thermal exchange with cold bodies. However, in the case of oxygen, nitrogen and hydrogen, this is not possible since the critical temperatures are very low. In order to liquefy these gases, we must resort to Joule-Thomson or adiabatic cooling.

In the former case, the gas is compressed by means of a compressor and passed through a refrigerator. Then, the gas enters a spiral tube and is allowed to escape through an aperture, which serves as the porous plug (partition) in the Joule-Thomson experiment (see p. 130), to a region of lower pressure (atmospheric). Upon expanding, the gas cools, rises to the top, and cools the spiral tube. Thus, each successive portion of escaping gas will be colder than the preceding one. Finally, a temperature is reached at which the gas is transformed into liquid.

The other method of liquefying gas involves the use of an expander. In a reciprocating expander, a gas is expanded adiabatically, performing work in moving a piston, and leaves the cylinder at a lower temperature. By having each portion of gas cool the subsequent one, one can reduce the temperature to  $-150^{\circ}\text{C}$ . Further cooling is impeded by the absence of suitable lubricants to maintain the friction between the piston and the cylinder walls at a low level. A solution to this problem was found by P. L. Kapitsa, who developed a refrigerating turbine—a turboexpander. A turbine is rotated by means of the gas from a compressor. The gas expands adiabatically, is cooled, and cools the subsequent portion of gas. Difficulties of lubrication are overcome by placing the bearings, requiring lubrication, external to the cold region.

Sec. 261. GAS  $\rightleftharpoons$  CRYSTAL TRANSFORMATIONS

When it is said that a substance "vaporises", the reference is usually made to the vaporisation of a liquid. The vaporisation of solids is called sublimation. One of the most familiar examples of evaporation of a solid is the sublimation of naphthalene.

Every odorous solid sublimates to a significant extent. The odour is produced by the molecules separating from the substance and reaching our olfactory organs. Usually, however, a substance will sublime to an insignificant extent. Sometimes, sublimation may not be detected even by very careful investigation. But, in principle, all solids, including iron and copper, vaporise. If sublimation is not detected it simply means that the density of the saturated vapour is extremely low. That this should be so is quite natural. The motion of atoms and molecules in a solid is very ordered and there is little probability that a molecule will separate from the surface of the solid.

The density of a saturated vapour in equilibrium with a solid increases with increasing temperature. It can be shown that a number of substances having a strong odour at room temperature do not manifest it at a reduced temperature. In most cases, the density of the saturated vapour of a solid cannot be increased significantly simply because the substance melts first.

Vapours are frequently used to obtain crystals when the latter are required in very pure form. This can be accomplished, for example, by precipitation on slightly cooled glass.

Sec. 262. LIQUID  $\rightleftharpoons$  CRYSTAL TRANSFORMATIONS

The transition from a liquid to a solid state (crystallisation) and the reverse transition (fusion) involve a fundamental rearrangement of particles. Upon fusion, long-range order in the arrangement of molecules or atoms disappears.

For a given pressure, fusion occurs at a very definite temperature. The vibrations of molecules or atoms become so intense that the maintenance of long-range order becomes impossible.

If the temperature is maintained at the fusion point, a liquid and a crystal may remain in a state of equilibrium, like in the case of a liquid and a saturated vapour. Crystals will neither grow nor melt.

External pressure changes the fusion temperature. As a rule, the fusion temperature increases with pressure, i.e., fusion becomes more difficult. However, there are several exceptions to this rule. One of these is ice. The melting of ice is facilitated by an increase in pressure. In terms of a phase diagram, we can briefly describe the normal and anomalous behaviour of bodies as follows: Usually,  $\frac{dp}{dT} > 0$ , i.e., the equilibrium curve forms an acute angle with the temperature axis. In the anomalous case,  $\frac{dp}{dT} < 0$  and the curve forms an obtuse angle with the abscissa. The strange behaviour of ice is related to another anomaly, namely, ice is lighter than water. The relationship of these two anomalies is shown in the following equation, which was derived above:  $\frac{dp}{dT} = \frac{Q}{T(v_1 - v_2)}$ . The overwhelming majority of solids are denser than their liquids. Evidently, under such conditions, a pressure which produces packing should facilitate fusion. The relationship between the two anomalies is quite natural. Consider a liquid and a crystal in a state of phase equilibrium. Let us raise the pressure without changing the temperature.

The atoms should approach one another. If the solid is denser, the liquid is transformed into crystalline state. If the liquid is denser, the reverse transition occurs.

The anomalies of water play an extremely important role in our lives. If they did not exist, rivers would freeze at the bottom. The anomalies of ice and other such bodies are due to the structures of these bodies. Ice crystals do not conform to the law of compact packing of particles. Thus, a disturbance of long-range order results in an increase in density rather than a decrease, as is usually the case.

Let us return to Fig. 275 (see p. 471). The broad ice channels may displace molecules of water by expanding somewhat. When ice melts, molecules may "fall" into this channel. In such a case, of course, the density will increase. No theory exists by means of which the heat of melting or the melting temperature could be predicted. This is due to the dependence on a great many structural factors. To be sure, the heat of melting is easier to predict, since the melting temperature is equal to the heat of melting divided by the entropy of melting.

A measure of the binding forces between molecules or atoms is, of course, the heat of sublimation (the energy required to break the intermolecular bonds) rather than the heat of melting (the energy required to eliminate long-range order).

As was indicated above, fusion of crystals cannot be retarded. On the other hand, crystallisation can be retarded and, in fact, sometimes will not occur at all. In order for crystallisation to begin in a liquid, there must appear a nucleus, i.e., a system consisting of several scores of atoms or molecules which have assumed an arrangement corresponding to that of a crystal of the substance. Moreover, conditions in the liquid must be favourable for the growth of this nucleus. In most liquids, it is difficult to achieve a significant retardation in the process of formation of nuclei. Cooling under strict conditions is required to achieve such retardation. Dust particles must be prevented from falling into the liquid and all mechanical disturbances such as vibrations and the jarring of the vessel containing the liquid must be avoided.

At sufficiently great supersaturation, it is probably impossible to avoid the spontaneous formation of nuclei, i.e., to avoid the stabilisation requisite for the crystallisation of atomic or molecular groups. But something else may occur. When the temperature is decreased, the mobility

of the particles may decrease to such an extent that the rate of growth of crystalline nuclei approaches zero. That is how glass is formed. Crystals will invariably grow when a few crystal particles (seeds) are to be found in a liquid under conditions of thermodynamic equilibrium with a crystal phase. Crystals for industrial purposes are grown by means of seeding.

If heat is removed very slowly, i.e., the temperature decreases a fraction of a degree per day, and if the crystalline particle turns in the liquid, the crystal grows only on a few of the faces possible. The growth occurs, of course, on the faces having the least surface energy. The indexes of such faces are always prime numbers. Most frequently, faces with the highest surface density of atoms or

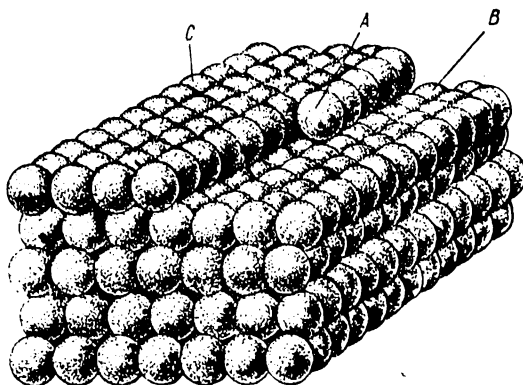


Fig. 295

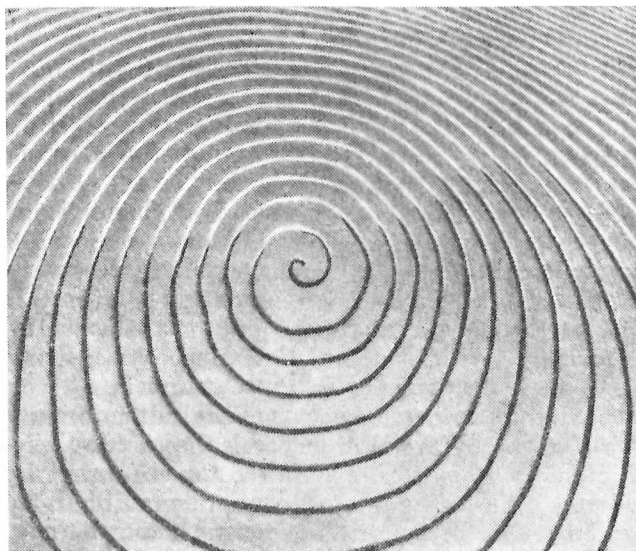


Fig. 296

molecules are formed. It is difficult to determine beforehand which faces will grow, especially since this depends on numerous subsidiary circumstances. Moreover, we shall not consider the growth of crystals from solutions having certain peculiar features. However, it may be asserted that a crystal in equilibrium with a melt (or a solution) will assume a form such that its surface energy is a minimum.

The mechanism of crystal growth consists in each successive particle attaching itself to the crystal at a point where the binding forces are a maximum and, therefore, the potential energy a minimum. Figure 295 shows three possible phases at which an atom can become attached to a growing crystal. At *A* the attracting forces acting on an atom are greater than at *B*, and at *B* greater than at *C*. Thus, a molecule or atom will invariably attach itself more easily to a layer already partially formed rather than begin to form a new layer.

According to calculations, in certain cases the initial formation of a new layer is bound up with such difficulties that the very growth of a crystal becomes incomprehensible. In such a case, the spiral mechanism of explaining growth is most applicable. It is evident from Fig. 283 that spiral growth can continue indefinitely and new atoms and molecules continually become attached to points which are favourable from the energy viewpoint. Thus, it is not necessary for growth to take place through the formation of a new layer. The beginning of spiral growth takes place with the formation of a fault known as a *spiral dislocation*. Such a "fault" is usually due to a tiny foreign inclusion. A portion of the surface of a crystal which has developed spirally is shown in Fig. 296.

#### Sec. 263. CRYSTAL $\rightleftharpoons$ CRYSTAL TRANSFORMATIONS

A transformation in the solid phase consists in a transition from one long-range order to another. The mechanism of such transformations is of great interest.

The simplest picture for the transformation of one solid phase into another in the case of simple substances is obtained when the structure of both phases consti-

tutes compact packings of spheres. Thus, cobalt and thallium are encountered in the form of cubic as well as hexagonal packings. By shifting a layer, we can transfer it from a "hexagonal" to a "cubic" state, and vice versa.

It even has been possible to grow a single crystal of a hexagonal phase from a single crystal of a cubic phase by means of transformations of this kind. Usually, this is not possible since the growth of crystals of a new phase begins simultaneously from many centres and a monocrystal becomes transformed into a fine crystalline substance. In most cases, a crystal crumbles when it is transformed into another solid phase. Sometimes the outer "shell" of a polyhedral monocrystal is preserved and a fine crystalline substance occupies this perfectly symmetrical volume.

The reason for the difficulty is clear. Crystals of a new phase may begin to grow from various points. But close layers in a cubic face-centred lattice may be formed

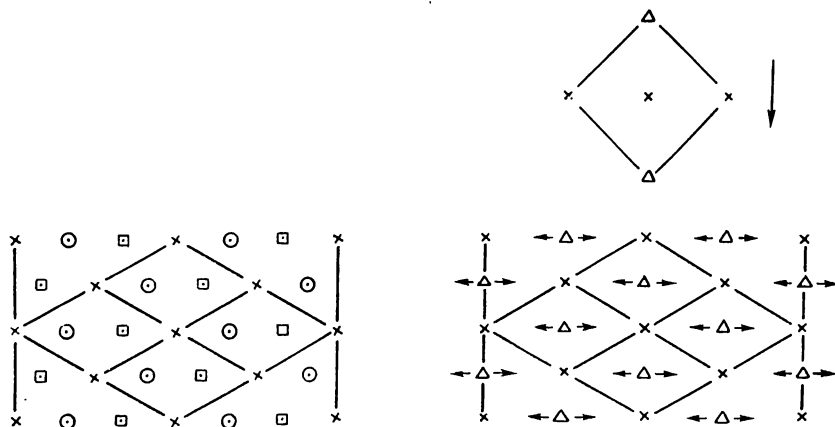


Fig. 297

by four different systems. Let us return to Fig. 272 (p. 475). In the crystal shown in this figure, close planes are perpendicular to the spatial diagonals, of which there are four in a cube (eight corners). Thus, hexagonal crystals of four different orientations may grow from a crystal with a cubic packing arrangement. G. V. Kurdjumov's work, devoted to transformations of iron and steel, laid the basis for the study of the rearrangement of atoms in phase transformations. At high temperatures, iron exists in the form of a compact cubic packing of atoms. At low temperatures, the iron atoms become arranged in a body-centred lattice. This transformation, known as a martensite transformation, is of tremendous importance in metallurgy\* and, therefore, should be considered in greater detail.

In Fig. 297, we see what occurs when the temperature is increased. The left diagram again shows a compact cubic packing arrangement; the right diagram shows a body-centred packing arrangement drawn in a rather unusual form: we see the projection of an arrangement of atoms as it appears when viewed along a plane diagonal of a cube. It would seem that these two diagrams have little in common, the main difference being that the left diagram represents a three-storeyed structure and the right a two-storeyed structure (the triangles are in the second storey). Less important is the difference in the angles of the rhombuses

\* The hardening of steel is nothing more than a martensite transformation.

(not shown in the figure). As the temperature is increased, the atomic vibrations increase and the less compact body-centred lattice becomes less advantageous at a temperature of  $906^{\circ}\text{C}$ . The two-storeyed structure becomes transformed into the three-storeyed one by the alternate shifting of the layers marked by triangles. For example, the odd layers shift to the left and the even ones to the right. This shift occurs along the diagonal of a rhombus; the angle of the rhombus changes at the same time.

When a phase transformation occurs in iron, crystal particles of the new phase may become oriented in any of 24 different directions. The number 24 is obtained in the following manner. There are four close planes in a cubic face-centred crystal, and, as can be easily shown, a crystal of the new phase grows in six different directions in a close layer.

Undoubtedly ordered, regular processes play an important role in the transition from one order to another. In such a rearrangement of order, atoms do not have to interchange places, i.e., only an organised shifting of atoms occurs. This is what takes place in a martensite transformation, a transformation which does not involve diffusion. However, in other transformations in a solid, diffusion phenomena may play an important role.

#### Sec. 264. DIFFUSION IN SOLIDS

It has long been known that foreign atoms diffuse in a solid. The surface layer of steel may be impregnated with carbon (cementation), nitrogen or boron. Diffusion occurs to a great depth and it is not particularly difficult to follow the process. At a temperature of  $200\text{--}300^{\circ}\text{C}$  significant quantities of silver penetrate lead to a depth of several centimetres in an hour.

However, not only foreign atoms migrate in a crystal. An iron atom can migrate in a crystal of iron and a copper atom in a crystal of copper. If a piece of radioactive copper is pressed against an ordinary piece of copper, the latter will soon become "contaminated" (radioactive). By means of tagged atoms, one can study the diffusion of atoms of the same kind, as well as of "foreign" atoms.

Diffusion is possible as a result of thermal vibrations. When an atom leaves its equilibrium position, a neighbour takes its place. Upon returning, the atom occupies the vacated spot. Thus, atoms interchange positions. Such an interchange is, of course, not easily achieved if only two atoms participate in the process. When two atoms interchange positions in a solid, a whole group of atoms are involved. An atom slips forward only when the thermal vibrations of many atoms accidentally create favourable conditions for this.

Dislocations, empty spaces and fractures, which always exist in a crystal, play an important role in diffusion. The presence of an empty space in a crystal facilitates the step-by-step migration of an atom through the lattice. An atom hindering this migration is "pushed" into the empty space.

If a foreign atom is not very large, it may move through the lattice without interchanging positions with lattice atoms. When the conditions become favourable, such an atom slides from one empty space in the compact packing of spheres to the next.

Diffusion is a two-way effect. If a zinc plate is pressed against a copper one, zinc atoms will penetrate the copper and copper atoms will penetrate the zinc. To be sure, the rates of flow in the two directions may vary considerably.

The diffusion of atoms through a crystal depends on many factors. It is interesting that a diffusion process proceeds most rapidly when the foreign atoms differ in all respects from the atoms of the crystal through which they move. Diffusion



proceeds most slowly for atoms which are the same as those of the crystal or in the same column of the Mendeleev periodic table as those of the crystal.

As already indicated, the presence of fractures and dislocations facilitates diffusion. Therefore, diffusion proceeds most rapidly in a deformed metal. The rate of diffusion is greatly dependent on temperature. This is not surprising since the diffusion coefficient, the coefficient of proportionality between flow of matter and concentration gradient, can always be represented by an expression of the form

$$Ae^{-U/kT},$$

where  $U$  is the height of the potential barrier which an atom must surmount in an elementary diffusion event. The existence of such a relationship is rather evident since the diffusion coefficient must be proportional to the number of atoms the energy of which suffices to cross the potential barrier.

Such barriers are quite high. In the case of self-diffusion, they are usually about 1-2 eV. It will be recalled that  $kT$  at room temperature is equal to  $\sim 0.03$  eV. The number of atoms with energies much greater than the average is very small; hence, practically no diffusion occurs. At a temperature of the order of  $1,000^\circ\text{C}$ , the situation is entirely different.

In the preceding article, we discussed crystal-crystal transformations occurring in an organised manner without diffusion. It should not be assumed that all phase transitions occur this way. On the contrary, at a sufficiently high temperature, the interchange of positions by atoms begins to play a very important role and the organised nature of transitions will be achieved in only small regions, or may even be completely obscured by the interchange of positions by atoms.

If a solid state transformation is of a diffusive nature, it proceeds at a rate that is comparable to that of self-diffusion processes. The heights of potential barriers surmounted by atoms during rearrangement are of the same order as during self-diffusion processes.

In the case of organised displacements of atoms of the martensite transformation type, the transformation proceeds at a ten-fold rate at low temperatures.

## Deformations of Bodies

### Sec. 265. ELASTIC PROPERTIES

For every solid, there exists a distorting force limit up to which a deformation is elastic. This means that if the elastic limit is not exceeded, the body returns to its original state.

Elastic deformations, like other deformations, are associated with the displacement of atoms (or molecules). When a body is extended elastically the interatomic spacing increases and when it is compressed the interatomic spacing decreases.

The distinctive feature of elastic deformations is that they do not destroy interatomic bonds or create new ones.

When a crystal is deformed elastically, all of the atoms continue to have the same neighbours. Thus, in elastic displacement, the lattice of a crystal as a whole becomes deformed (sloped). Hence, each atom continues to have the same neighbours. This enables the body to return to its equilibrium state when the distorting force is removed.

The change in interatomic spacing that may be achieved by means of elastic extension or compression is quite small. The maximum relative elongation of an elastic nature does not, as a rule, exceed 0.001. This means that for interatomic spacings of the order of  $2\text{\AA}$  the equilibrium positions of the atoms may be displaced by no more than  $0.002\text{\AA}$ . Such small changes in a lattice period can be detected through X-ray analysis by observing the displacement of diffraction lines on a roentgenogram. This requires that the lines be filmed at large  $\theta$  angles, since only in this manner can small changes in interplanar spacings be detected (see p. 297).

The elastic deformation of polymers such as rubber is of an entirely different nature. Rubber has mechanical properties which basically differ from those of crystalline substances. The fundamental difference lies in the magnitude of elastic elongation. Certain kinds of rubber may be stretched to 10-15 times their normal length without exceeding the elastic limit. Thus, they may be elongated 10,000 times more than metals! The magnitude of the modulus of elasticity of rubber is no less striking.

A steel wire having a cross-section of  $1\text{ mm}^2$  will stretch one twenty-thousandth of its length under the action of a load of 1 kgf, but a rubber band having the same cross-section will stretch to twice its original length under the action of such a load.

Two processes occur when polymers are stretched. First, tangled bundles of molecules become disentangled. At the same time, there occurs regular packing of certain portions of the disentangled bundles of molecular chains into a three-dimensional order. Evidently, the fact that crystallisation takes place upon stretching is of secondary importance, since the established order does not remain when the external force is removed, i.e., the bundles of molecules become twisted once again.

The twisting of bundles of molecules is accompanied by an increase in entropy, i.e., an increase in the degree of disorder. It turns out that the internal energy of rubber and similar polymers practically does not change when such substances are elastically deformed. Therefore, the work of stretching, which according to the fundamental laws of thermodynamics is  $dA = dU - T dS$ , in this case simply

equals  $-T dS$ , i.e., it is directly proportional to temperature. (It should be recalled that the work of external forces on a system is considered to be negative.) In this respect, the elastic deformation of rubber is of the same nature as the isothermal compression of a gas (cf. p. 132).

It should be noted that both the elastic deformation of a crystal and that of rubber do not yield a new potential energy minimum. In the case of a crystal, this is due to the fact that we do not go out of the potential well, and in the case of rubber—that the energy does not change at all.

## Sec. 266. PLASTIC PROPERTIES

**Slippage.** The elastic deformation of a crystal consists in the changing of its interatomic spacing with each atom maintaining its same neighbours. On the other hand, in the case of a plastic deformation—a deformation, which remains when the external force producing it is removed—atoms surmount their potential barriers and enter new “potential wells”, i.e., change their neighbours. The basic mechanism of plastic deformation is the slippage of one atomic plane relative to another. An element of such slippage consists in the displacement of all of the atoms by one period. This can be detected with the naked eye in the form of so-called slip bands. Slippage occurs at the weakest points (fractures and other defects) and the crystal breaks up into layers (slip stacks). The plotting of stack thickness, the order of magnitude of which is equal to several tenths of a micron, yields a random distribution curve. The forces required to displace atomic planes having different indexes will differ. Usually, it is easiest to displace the planes which are most compactly filled with atoms. However, slip planes of crystals may change with changes in temperature and impurities and also during the deformation process itself. In aluminium, the plane (111) is a slip plane.

Slippage occurs along a given plane having a definite orientation. Usually this plane has the densest distribution of atoms (e.g., [101] in an all-sided face-centred cubic lattice).

In order for displacement to begin, a certain minimum stress (ultimate shearing stress) is required. The magnitude of this stress is very small, reaching several grams per square millimetre in some cases. In measuring the ultimate stress, one must, of course, take into account the orientation of the slip plane relative to the external force.

**Strength.** A monocrystal of zinc can be easily bent by hand. However, it is not possible to straighten it in the same manner. This is due to the fact that its strength has increased.

The shearing strength of a crystal increases with increasing deformation. Therefore, plastic displacement along a given slip plane does not cause the material to rupture, but rather ceases when the strength is sufficient to oppose the external force, and thereupon displacement begins in other planes. Thus, the number of slip bands increases and the slip stack thicknesses decrease.

If an external force is applied to a crystal previously subjected to plastic deformation, such deformation will resume, of course, when the magnitude of the force reaches the value at which it previously ceased to be effective owing to an increase in strength. It can be stated, therefore, that an increase in the strength of a crystal increases its elastic limit—and, moreover, by a large factor.

One theory relates the described increase in the strength of a crystal to a disturbance of the regularity, i.e., distortion, of its lattice. From this viewpoint, it is quite natural that the strength of a crystal should increase with increasing rate

of deformation and decrease with increasing temperature. However, from the viewpoint of the dislocation theory discussed above, the picture is different.

**Plastic Deformation as a Displacement of Dislocation.** Let us consider in greater detail the process of displacing one atomic plane relative to another. If there are no dislocations in a slip band, it is necessary to shift every row of atoms in the displacement plane. The situation is quite different when a shearing force acts on a crystal containing dislocations.

Figure 298 shows a compact packing of spheres which contains a simple dislocation (only the end spheres of the rows are shown). For simplicity, let us assume that the dislocation region embraces a minimum number of rows. Then, the dislocation consists basically in the following: between two rows of the upper, extended layer, adjoining the boundary between blocks, there is a linear gap. In the

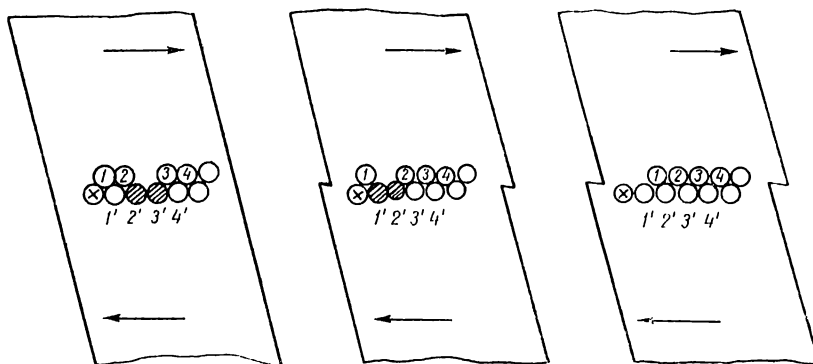


Fig. 298

lower, compressed layer on the other side of the boundary between blocks, there is an extra row of atoms (the two rows of atoms just above the linear gap are very compressed). Now, let us begin to displace the upper block to the right relative to the lower one. At a certain initial instant, the "fissure" is between rows 2 and 3, rows 2' and 3' are compressed. As soon as the force becomes effective, row 2 moves into the "fissure", the shape of sphere 3' is restored and sphere 1' becomes compressed. The entire dislocation has shifted to the left, and it will continue to move in this direction until it is "pushed out" of the crystal. In other words, displacement consists in shifting the dislocation line along the displacement plane. It is clear that a much smaller force is required to achieve displacement when dislocations are present.

According to calculations, the strength of a crystal in which there are no dislocations should be a hundred times as great as the strength of an actual crystal, determined experimentally. The presence of a small number of dislocations suffices to decrease the strength to a small fraction of that of an ideal crystal.

Figure 298 shows how a dislocation is "pushed out" of a crystal by an applied force. Thus, as the degree of deformation is increased, the strength of a crystal increases. When the last dislocation of a crystal is eliminated, its strength will be about a hundred times as great as that of a perfectly normal crystal. In this manner, an increase in strength can be easily explained. To be sure, in order to obtain quantitative agreement between calculations and experimental results, we must assume that helical as well as ordinary dislocations may aid displacement.

Excellent confirmation of this theory is provided by the fact that the strength of perfect crystals which are grown artificially is approximately equal to the calculated value for an ideal crystal.

#### Sec. 267. ULTIMATE STRENGTH

As the stress within a body increases, its deformation increases up to a certain point (the elastic limit). Then, plastic deformation begins and because of an increase in strength the curve rises sharply. Finally, the rupture point is reached.

The ultimate strength of a body depends, to a certain extent, on the duration of an applied force. When the force is of prolonged duration, the ultimate strength drops. The values given in engineering handbooks generally refer to short-duration tests. The dependence of the ultimate strength of a body on the duration of an applied force indicates that in slow processes the behaviour of a solid depends on internal diffusion processes. It may be assumed, for example, that even small forces can create a favourable trend for processes involving an interchange of positions by atoms.

The value calculated for the ultimate strength of a perfect crystal is several hundred times as great as measured values. This signifies that faults play a fundamental role in a crystal. The calculated value for the rupture strength of a monocrystal of rock salt is  $200 \text{ kgf/mm}^2$ . Under ordinary conditions, a rod of this substance will rupture when it is subjected to a load of  $0.5 \text{ kgf/mm}^2$ .

The role played by fractures was demonstrated in the well-known experiments of A. F. Yoffe, who investigated the rupture of rods of rock salt in water. Water dissolves the surface of such a rod and "heals" the microfractures formed during extension. As a result, the rupture strength of rock salt becomes greater than  $100 \text{ kgf/mm}^2$ , i.e., it approaches the theoretical value.

The presence of faults serves to decrease the effective area at which rupture occurs. In the final analysis, the force acting at the instant of rupture is determined by the number of broken interatomic bonds.

The strength of various materials determined in this manner vary within narrow limits. Thus, a piece of thread or rubber band ruptures at a stress of the same order of magnitude as that of a steel wire.

#### Sec. 268. MECHANICAL PROPERTIES OF A POLYCRYSTALLINE MATERIAL

During the initial, elastic stage of deformation, the grains of a crystalline material are not deformed uniformly since they have various orientations relative to the line of action of the force. As a result, the elastic properties of a polycrystalline substance will differ from those of a monocrystal.

However, in the plastic region, the behaviour of a polycrystalline substance may differ from the behaviour of a monocrystal of the same substance to an even greater extent. A polycrystalline substance offers greater resistance to an external force. This is not surprising since the development of plastic displacements in a given grain will be impeded by neighbouring grains in which slip planes are oriented entirely differently relative to the applied force.

Moreover, in a polycrystalline material, there occurs an entirely new phenomenon: the turning of grains and the formation of texture. The turning of grains in drawing, rolling and other deformation processes is determined by the tendency of each grain to become aligned in such a manner that slippage is facilitated, i.e., with its slip plane parallel to the applied force. If there are several slip planes, the grain assumes a position for which the effect of the slip planes is maximal.

For example, in most metals having cubic all-sided face-centred cells, the grains tend to become arranged with the [111] directions parallel to the axis along which the material is drawn.

Another new phenomenon occurring in such a material consists in the slippage of the grains relative to one another along an intercrystal layer. Such displacement differs from crystal displacement and is more like viscous flow in thick liquids.

A polycrystalline material may not rupture at the same value of stress as a monocrystal. In certain cases, the grains of a polycrystalline material remain whole when the material ruptures. This occurs when intercrystal layers have weak mechanical properties. As a general rule, the strength of a material increases with decreasing grain size.

#### Sec. 269. THE EFFECT OF SURFACE-ACTIVE SUBSTANCES ON DEFORMATION

Every solid has numerous ultramicroscopic structural faults, arising as the result of the thermal mobility of its atoms and the presence of impurities and mechanical defects. These faults are distributed throughout the volume of a solid and in many cases can be viewed as extremely minute embryonic microfissures. The basic characteristic of such embryonic microfissures is their ability to increase in size during the process of deformation of the material.

Under the action of external forces, the size of these microfissures increases and a concentration of stress arises at their edges. This, in turn, further facilitates the growth of the microfissures. In the case of a brittle material, such an increase in the size of microfissures during a deformation process may result in premature rupture of the material. In the case of a plastic material (most metals) such an increase in the size of microfissures during a deformation process results in the formation of plastic displacements. When a body is in a three-dimensional stressed state, the microfissures have wedge-shaped cross-sections and are characterised by exposed surfaces—orifices and cul-de-sacs in which the fissures preserve their embryonic nature. Actual fissures terminate in cul-de-sacs like a sharp blade having a very large curvature, with a radius of curvature of the order of magnitude of the lattice spacing. When the deforming forces are removed—in the range of elastic deformations—the microfissures gradually “heal” in reverse order, i.e., first the cul-de-sacs close and then the orifices.

P. A. Rebinder has shown that the effect of the surrounding medium on mechanical properties of a solid is not restricted to such chemical action as corrosion and dissolution. The exposed surface of a solid is always coated with a thin film of a component of the surrounding medium which is most closely related to the given solid. Such a substance may be a gas or vapour usually contained in air, or a substance located in the vicinity of the solid. The molecules of a substance clinging to the surface of a solid, or, as we generally say, adsorbed by a solid, are able to move along this surface and migrate from a region where there is an excess of such molecules to a region where there is a deficiency for complete coating of the surface. The tendency of an adsorbed layer to occupy all of the surface available to it is due to the fact that adsorption decreases the surface energy of a solid. Substances which may be adsorbed by the surface of a solid are called surface-active substances. Various organic alcohols, acids, and salts of these acids, i.e., soaps, are highly surface-active substances with respect to metals.

The strength of a solid decreases when it adsorbs a surface-active substance. If a solid is ruptured in a medium containing even a small amount of a surface active substance (for example, in a solution of oleinic acid in pure vaseline oil), the force required to rupture the material is less than under ordinary conditions.

This is particularly evident in the crushing of soft rocks and in the subjection of metals to a variable force or a force of prolonged duration.

The effect of adsorbed molecules on the strength of a solid can be explained as follows. When molecules are adsorbed by the surface of a body, they penetrate the microfissures as a consequence of their mobility and tendency to occupy all of the exposed surface of the adsorbent. The drawing of adsorbed layers into a microfissure is due to the decrease in the surface energy of a solid caused by such penetration. If an obstacle is placed in the path of an adsorption layer tending to occupy a surface area of a solid which is not yet occupied, the adsorption layer will exert pressure on the obstacle. Within a microfissure, such an obstacle is provided by the molecules themselves, i.e., their size prevents them from penetrating deeper into the microfissure. Therefore, at the boundary of an adsorbed layer within a microfissure, a pressure, which is directed so as to increase the size of the fissure in depth, arises. Adsorbed layers behave like wedges driven into the microfissures. Thus the penetration of adsorbing molecules into the orifices of microfissures tends to create additional disrupting forces. This is equivalent to increasing the external deforming forces. Therefore, the rupture of a solid in the presence of adsorbing substances is brought about by lower applied forces.

#### Sec. 270. MATERIAL BREAKDOWN UNDER THE ACTION OF A STREAM OF PARTICLES

The problem of material breakdown under the action of a stream of particles is of great importance in the construction of nuclear reactors. The materials of a reactor, including nuclear fuel, moderator, walls and instruments, are subjected to the action of neutrons, fission fragments, electrons, etc. Let us consider those actions of streams of particles which leave permanent effects.

In the first place, we should mention particle collisions in which an electron providing a chemical bond between atoms is dislodged from its position. In such cases, ionisation results in the rupture of the bond. This bond is not necessarily re-established. Moreover, the ions or radicals which are formed may recombine in a different way. Therefore, in molecular materials, ionisation results in the breakdown of certain molecules and the creation of new ones.

An atomic nucleus may also be displaced from its position. In such a case, it drags along its electron shell. Therefore, it can be said that an entire atom, rather than just a nucleus, has been dislodged from its position. Such an effect of the action of radiation is almost always irreversible.

Materials are damaged by radiation as a result of the displacement of atoms from their positions and the rupture of chemical bonds. Atoms are dislodged under the action of heavy charged particles and fast neutrons. Chemical bonds are ruptured under the action of slow neutrons,  $\gamma$ -rays and electrons.

Let us consider more carefully what occurs when atoms are dislodged from their positions in solids. The process of displacement of atoms is a chain process. This means that the first displaced atom displaces another atom which is located in its path; the latter is able to displace a third atom, etc. By means of such a chain process, a single fast nuclear projectile is able to produce considerable distortion in the crystal lattice of a solid. The nature of the distortion varies considerably from case to case. A crystal lattice may be completely destroyed. Foreign atoms may penetrate between atoms of the primary lattice. Also possible are processes involving the substitution of atoms of the primary lattice by projectile atoms. The number of displacements per charged particle for one element does not differ greatly from the number for another element. An  $\alpha$ -particle having an energy of 5 MeV, or a proton having an energy of 20 MeV, dislodges 60-80 particles.

These figures suggest that the heavier the particle, the greater the damage. Analogous figures for fission fragments of uranium-235 or plutonium-239 nuclei are considerably more unexpected. Such a pair of fragments produces, for example, 25,000 displaced atoms in uranium and 8,000 displaced atoms in graphite.

The process of slowing down a neutron from its initial velocity to thermal velocity also does not proceed without causing damage to materials. The slowing down of a neutron causes 450 atoms to be displaced in beryllium, 1,900 atoms in graphite and 6,000 atoms in aluminium.

It is evident from these figures that significant changes in the properties of a crystal lattice are to be expected as a result of the displacement of atoms from their positions in the lattice. Metals are of primary interest in this connection. The reason for this is that the sole permanent effect of radiation in metals is the displacement of atoms.

The action of neutrons and fission fragments has been studied most carefully. This is not surprising since such investigations are of basic importance in the design of nuclear reactors. The effect of a dose of  $10^{19}$  neutrons per sq cm has been studied in detail. Such a dose is not very large. Generally speaking, the materials of a nuclear reactor are subjected to such a stream of neutrons during each day of operation. However, even in the case of such a small dose, the properties of metals undergo important changes. These changes approximate those which occur in the cold working of metal. Thus, under the action of neutrons and fission fragments a metal's brittleness and hardness increase, its ductility decreases, and its electromagnetic properties also change.

Radiation damage in metals consists mainly in the displacement of atoms from their positions, but in organic materials, where the atoms are connected by chemical bonds, the changes consist mainly in the rupture of such bonds as the result of ionisation. Organic materials are very rapidly broken down under the action of radiation. A dose of the order of  $10^{19}$  neutrons per square centimetre practically disintegrates an organic substance. In a reactor, paraffin, olefin and polyphenyl sustain a damage of 25 per cent within several hours.

The transformation of an organic material usually consists in the liberation of a gas and in polymerisation. However, it should be noted that in certain cases high-polymer materials are depolymerised under the action of radiation.

The crystal lattice is arranged so that we may expect different easiness of penetration of a charged particle in different crystallographic directions.

At first, it may seem that it is rather difficult for a charged particle to get into a lattice tunnel. It is easy to compute that for the thinnest crystals the motion of a proton along a straight line without "brushing" against the atoms located in its way must be realised with an accuracy to 0.01 degree of arc. However, the experiments carried out recently on monocrystals showed that tunnelling can be observed without special difficulties. When the direction of a beam of particles coincides with the axes of crystals, the intensity of proton current increases five to ten times. A detailed investigation showed that tunnelling is a peculiar process with a feedback. When a proton deviates from its path along an atom row, the electrostatic forces return it back to the straight path.



# Dielectrics

## Sec. 271. THE RELATIONSHIP BETWEEN PERMITTIVITY AND THE POLARISABILITY OF A MOLECULE

In a number of cases, particularly in gases, the molecules of a substance do not interact with one another. Hence, the electrical properties of such a substance are determined by the average behaviour of one of its molecules. Molecules do not interact with one another in many dilute solutions as well. Occasionally, molecular interaction plays a secondary role even in a condensed phase.

Therefore, consideration of the electrical properties of a substance which consists of a large number of noninteracting molecules is of considerable importance.

The dipole moment of a unit volume of dielectric  $P$ , is determined by the permittivity  $\epsilon$  and the field intensity  $E$  in accordance with the formula

$$P = \frac{\epsilon - 1}{4\pi} E$$

(see p. 193). On the other hand, the polarisation vector  $P$  is equal to the sum of the dipole moments in a unit volume of dielectric:

$$P = \sum p$$

or

$$P = N\bar{p},$$

where  $N$  is the number of molecules in such a unit volume and  $\bar{p}$  is the "contribution" of each molecule to the polarisation vector. If  $E'$  is the field intensity which acts on a molecule, then

$$p = \beta E',$$

where  $\beta$  is the polarisability of the molecule.

It would seem that the relationship between  $\beta$  and  $\epsilon$  should now be given by the expression  $\epsilon = 1 + 4\pi N\beta$ . However, this is not so, and that is why the field intensity in the above formula was denoted by  $E$  prime. The equations relating  $P$  and  $E$  and  $\bar{p}$  and  $E'$  involve different field intensities.  $E$  is the force acting on a unit test charge; such a charge does not distort the existing field.  $E'$  is the field produced by all the molecules (with the exception of the given one) acting upon the given molecule.

On p. 196, it was indicated that the field inside a dielectric sphere,  $E_i$ , is related to the external field in which this sphere is located as follows:

$$E_i = E_e - \frac{4}{3} \pi P.$$

It is evident that the field in a spherical cavity which is cut in a dielectric may be determined by changing the sign of  $P$ :

$$E_i = E_e + \frac{4}{3} \pi P.$$

It may be rigorously proved that  $E'$ , the field due to all the molecules of a gas (except the one acted upon), is equivalent to the field in a spherical cavity. Thus,

$$E' = E + \frac{4}{3} \pi P.$$

The relationship between  $\beta$  and  $\varepsilon$  may now be determined. Equating the derived expressions for  $P$ , we obtain

$$\frac{\varepsilon-1}{4\pi} E = N\bar{p} = N\beta E'.$$

Now, substituting  $E' = E + \frac{4}{3} \pi \frac{\varepsilon-1}{4\pi} E$ , we obtain the so-called *Clausius-Mosotti formula*:

$$\frac{\varepsilon-1}{\varepsilon+2} = \frac{4\pi}{3} N\beta.$$

If each member of the equation is multiplied by  $\frac{M}{\rho}$ , where  $M$  is the molecular weight and  $\rho$  is the density, the resulting expression will depend only on the polarisability  $\beta$ . Thus,  $N \frac{M}{\rho} = N_{Av} = 6.02 \times 10^{23}$  (Avogadro's number).

The quantity

$$\mathcal{P} \equiv \frac{\varepsilon-1}{\varepsilon+2} \frac{M}{\rho} = \frac{4\pi}{3} N_{Av}\beta$$

is called *the molecular polarisation*. To determine the molecular polarisation, one must first measure the permittivity of a substance as the ratio of the capacitance of a condenser filled with the substance under investigation to the capacitance of the condenser with the dielectric removed. Capacitance is usually measured by means of a bridge. Such bridges are constructed for the range 30 Hz to 300,000 Hz. However, bridges may be constructed for frequencies up to 40 MHz.

A wide variety of permittivity meters (instruments for measuring  $\varepsilon$ ) are employed. Such instruments are very sensitive and enable us to measure permittivity with a high degree of accuracy. Indeed, excellent results may even be obtained with gases having a pressure of the order of 1 mm of Hg.

Using the formula  $\varepsilon = n^2$  (see p. 250), we can derive an index of refraction equation which is analogous to the molecular polarisation equation:

$$R \equiv \frac{n^2-1}{n^2+2} \frac{M}{\rho} = \frac{4\pi}{3} N_{Av}\beta.$$

This characteristic of a molecule is called *molecular refraction*.

Measured values of  $R$  and  $\mathcal{P}$  for different frequencies of electromagnetic vibrations may differ considerably from one another.

Despite the fact that the derivation of these formulas presupposes a gas, molecular interaction evidently changes matters little. In any case, the formulas for  $R$  and  $\mathcal{P}$  are widely used in the investigation of dilute solutions as well.

*Examples.* Let us consider benzene,  $C_6H_6$  ( $\varepsilon = 2.28$ ;  $\rho = 0.88$  g/cm<sup>3</sup>;  $M = 78$ ), and water ( $\varepsilon = 81$ ;  $\rho = 1$  g/cm<sup>3</sup>;  $M = 18$ ). Assume that the plates of a flat condenser, which creates an electric field  $E = 300$  V/cm = 1 CGS unit, are immersed in these liquids.

1. Let us calculate the polarisation (the electric moment of a unit volume of dielectric) of benzene and water:

$$P_{benzene} = \frac{\varepsilon-1}{4\pi} E = \frac{2.28-1}{4 \times 3.14} \times 1 = 0.1 \text{ CGS unit};$$

$$P_{water} = 6.4 \text{ CGS units.}$$

The contribution of each molecule to the polarisation vector is  $\bar{p} = \frac{P}{N}$ , where  $N = \frac{N_{Av}\rho}{M}$  is the number of molecules per unit volume;

$$\bar{p}_{benzene} = 1.5 \times 10^{-23} \text{ CGS unit};$$

$$\bar{p}_{water} = 19.4 \times 10^{-23} \text{ CGS unit}$$

2. Let us calculate  $E'$ , the intensity of the electric field due to all molecules (except the one acted upon):

$$E''_{benzene} = E + \frac{4}{3} \pi P_{benzene} = 1.43 \text{ CGS units;}$$

$$E'_{water} = 27.8 \text{ CGS units,}$$

i.e., the field in water is about 28 times (1) as great as the applied field. Now, the polarisability of molecules of benzene and water can be determined:

$$\beta_{benzene} = \frac{\bar{p}}{E'} = 1.05 \times 10^{-23}, \quad \beta_{water} = 0.7 \times 10^{-23}.$$

3. From measurements of  $\epsilon$  by means of a permittivity meter, we can calculate the molecular polarisation  $\mathcal{P} = \frac{\epsilon - 1}{\epsilon + 2} \frac{M}{\rho}$ :

$$\mathcal{P}_{benzene} = 26.6 \text{ CGS units;}$$

$$\mathcal{P}_{water} = 17.3 \text{ CGS units.}$$

4. Measurements of the index of refraction  $n$  by means of a refractometer yield  $n_{benzene} = 1.5014$  and  $n_{water} = 1.330$ . Using these values, we can calculate the molecular refraction  $R = \frac{n^2 - 1}{n^2 + 2} \frac{M}{\rho}$ :

$$R_{benzene} = 26.1 \text{ CGS units;}$$

$$R_{water} = 3.6 \text{ CGS units.}$$

It is seen that in the case of benzene  $\mathcal{P} \approx R$  and in the case of water the values of  $\mathcal{P}$  and  $R$  differ considerably. The reason for this will be explained in the next article.

## Sec. 272. POLARISATION OF POLAR AND NONPOLAR MOLECULES

Polarisation of a substance under the action of an electric field may occur for two reasons. Firstly, the centre of gravity of the electron shell may be displaced (inherent polarisability). Secondly, the field has an orienting action which may turn molecules having a constant, or rigid, dipole moment closer to the direction of the field. Therefore, it is customary to divide polarisability into two parts:  $a$ —inherent polarisability and  $b$ —orientation polarisability.

A molecule must be turned as a whole in order for the dipole to become oriented. Owing to the inertia of a molecule, such turning requires a certain amount of time. For rapid electromagnetic vibrations, a rigid dipole cannot follow the field. Therefore, in the case of light waves, the orientation polarisability  $b$  is absent.

Thus,

$$\mathcal{P} = \frac{4\pi}{3} N_{Av} (a + b) \quad \text{and} \quad R = \frac{4\pi}{3} N_{Av} a.$$

The polarisability  $a$  of a molecule can be determined by measuring the index of refraction. If, in addition,  $\mathcal{P}$  is also measured, the orientation polarisability  $b$  is obtained by subtraction.

The magnitude of the orientation polarisability is directly related to the rigid dipole moment  $p$  of a molecule. We shall show that  $b = \frac{p^2}{3kT}$ .

Gas molecules are randomly oriented as a result of chaotic thermal motion. In the absence of a field, the assumption of any direction by the dipole moment  $p$  of a molecule is equally probable. The situation changes if a field  $E$  is applied. The potential energy of a dipole is equal to  $e(\varphi_+ - \varphi_-)$ , where  $\varphi_+$  and  $\varphi_-$  are the potentials of the field at the ends of the dipole, i.e.,

$$-e \frac{\partial \varphi}{\partial l} l = -pE = -pE \cos \theta,$$

where  $\theta$  is the angle between the field vectors and the dipole moment. A dipole oriented in the direction of the field has a minimum energy. This energy is equal to  $-pE$ . Thermal motion prevents all dipoles from assuming a position of minimum energy. A certain compromise distribution is established between the tendency to maximum entropy and the tendency to minimum energy (see p. 500). The Boltzmann law is an expression of this compromise. The probability that the energy of a molecule lies between  $U$  and  $U + dU$  is proportional to  $e^{-\frac{U}{kT}} dU$ . In our case,  $U = -pE \cos \theta$ . Therefore,  $dU = pE \sin \theta d\theta$ . The fraction of molecules the dipole moments of which lie between  $\theta$  and  $\theta + d\theta$  will be equal to  $e^{\frac{pE}{kT} \cos \theta} \sin \theta d\theta$ .

For ordinary temperatures,  $pE \ll kT$ . Even for extremely strong fields of the order of  $10^5$  V/cm, the ratio  $\frac{pE}{kT}$  will be of the order of 0.01 (the order of magnitude of dipole moments is  $10^{-18}$  CGS unit). Therefore, we can use the approximation  $e^x \approx 1 + x$ , and the fraction of molecules sought will be equal to

$$\text{const} \left( 1 + \frac{pE}{kT} \cos \theta \right) \sin \theta d\theta.$$

The integral of this expression from 0 to  $\pi$  should equal unity from the probability viewpoint, since for any molecule the direction of  $\mathbf{p}$  lies somewhere between 0 and  $\pi$ . Then, as can be easily verified, the constant is equal to  $\frac{1}{2}$  and the fraction of molecules the polarisation vectors of which lie in the interval from  $\theta$  to  $\theta + d\theta$  will be equal to

$$\frac{1}{2} \left( 1 + \frac{pE}{kT} \cos \theta \right) \sin \theta d\theta.$$

The projection of the dipole moment on the direction line of the field is  $p \cos \theta$ . If  $N$  is the number of molecules per unit volume, the fraction contributed to the polarisation vector by molecules inclined at an angle  $\theta$  to the field will be equal to

$$\frac{1}{2} N p \left( 1 + \frac{pE}{kT} \cos \theta \right) \sin \theta \cos \theta d\theta.$$

The polarisation vector  $P$  can be determined by integrating this expression from 0 to  $\pi$ . We obtain

$$P = N \frac{p^2}{3kT} E;$$

hence, the orientation polarisability is given by the formula

$$b = \frac{p^2}{3kT}.$$

The relationship between molecular polarisation and temperature is expressed by the formula

$$\mathcal{P} = \frac{4\pi}{3} N_{Av} \left( a + \frac{p^2}{3kT} \right).$$

This theoretical conclusion is in excellent agreement with experimental results. By measuring  $\mathcal{P}$  as a function of  $T$ , we can easily determine the two parameters which describe the electrical properties of a molecule, viz., polarisability and the "rigid" dipole moment  $p$ .

Thus, the values obtained for  $a$  by measuring  $R$  can be used to determine  $p$  by substituting it in the expression for  $\mathcal{P}$ .

Experiments indicate that in certain cases the interaction of dipoles of neighbouring particles may result in significant changes in permittivity as compared

with the value of  $\epsilon$  for a system of noninteracting molecules. This can be shown by measuring  $\epsilon$  in a liquid and in a gas formed of the same molecules.

The interaction of particles also affects the permittivity of crystals.

As a rule, electric polarisation in crystals occurs only as the result of the deformation of electron shells and the displacement of ions. No orientation polarisation occurs since, by and large, molecules cannot turn in a crystal.

In many ionic crystals, the index of refraction squared is considerably less than the permittivity. (For example, the values for rock salt are 2.37 and 6.3, respectively, for titanium dioxide 7.3 and 114, and for lead carbonate 4.34 and 24.) In such crystals, the electron shell is deformed and, in addition, the ions are displaced as a whole under the action of a static field. On the other hand, it has been established that in molecular crystals the permittivity is equal to the square of the index of refraction. This indicates that polarisation is due exclusively to the deformation of the electron shell.

Since orientation polarisation is absent, permittivity varies very little as the temperature changes.

It has already been indicated that in the case of a rapidly varying field there is no orientation polarisation and the molecular polarisation becomes equal to the refraction. It is important to know which field oscillations should be considered rapid. This can be determined if the relaxation time is known. When the relaxation time  $\tau$  is much greater than the oscillation period, there is no orientation polarisation.

The relaxation time  $\tau$  was discussed on earlier. If a dielectric is in a constant field, its dipoles assume an equilibrium orientation distribution which depends on the temperature. When the field is switched off, the dipoles become disoriented. However, this does not occur instantaneously, i.e., the order decreases in accordance with an exponential law. The rate of this decrease is described by the relaxation time  $\tau$ —the time in which the polarisation decreases to  $\frac{1}{e}$  of its original value. If  $\tau$  is much greater than the oscillation period, the direction of the external field changes before the dipoles change their orientation. A very rapidly varying field does not affect the behaviour of the dipoles at all. If  $\tau \ll T$ , at each instant the state will be in equilibrium and the polarisation will closely follow changes in the field. For most dielectrics, the relaxation time is of the order of  $10^{-12}$ – $10^{-13}$  sec.

*Examples.* 1. Using the results of the example on p. 520 let us determine the values of the inherent polarisability  $a$  and the orientation polarisability  $b$  for benzene and water:

$$a = 3R/4\pi N_{Av}, \quad \text{whence} \quad a_{\text{benzene}} = 10^{-23} \text{ CGS unit}; \\ a_{\text{water}} = 0.14 \times 10^{-23} \text{ CGS unit}.$$

On the other hand,

$$a + b = 3\mathcal{P}/4\pi N_{Av}, \quad (a + b)_{\text{benzene}} = 10^{-23} \text{ CGS unit}; \\ (a + b)_{\text{water}} = 0.7 \times 10^{-23} \text{ CGS unit}.$$

It follows that

$$b_{\text{benzene}} = 0 \quad \text{and} \quad b_{\text{water}} = 0.7 \times 10^{-23} - 0.14 \times 10^{-23} = 0.56 \times 10^{-23} \text{ CGS unit}.$$

This means that a benzene molecule does not have a rigid dipole moment, but a water molecule does.

2. Let us determine the rigid dipole moment of a water molecule from the formula  $p = \sqrt{3kTb}$ . If the molecular polarisation  $\mathcal{P}$  and the molecular refraction  $R$  are measured at room temperature ( $T = 300$  K),

$$p = \sqrt{3 \times 1.38 \times 10^{-16} \times 300 \times 0.5 \times 10^{-23}} = 0.8 \times 10^{-18} \text{ CGS unit}.$$

This value is in close agreement with experimental results.

Frequently, the unit 1 debye =  $10^{-18}$  CGS unit is used as a measure of dipole moment. This unit is named after the German scientist Debye, who developed the theory of dipole moments.

## Sec. 273. ADDITIVITY OF MOLECULAR REFRACTION

The refraction  $R$  is a molecular constant.  $R$  does not depend on the density or phase of a substance (this has been demonstrated experimentally), nor on the temperature of the given substance. A convenient property of refraction is its additivity. If it is possible to compile a table of property increments\* for all possible atoms, and if the magnitude of the property is determined as the sum of the increments, such a property is said to be additive. The additivity of  $R$  can be used for analytical and identification purposes. (It should be noted that this additivity possesses no theoretical basis and in a number of cases there occur significant deviations from the ideal.)

Innumerable observations have been processed by numerous investigators, and tables of  $R$  increments have been compiled. (Most of these tables are for  $R_D$  increments, i.e., measurements of the index of refraction for the so-called  $D$ -line, the yellow line of sodium.) For example, for C, H and Cl atoms, the increments are equal to 2.418, 1.100 and 5.967, respectively. By means of these values alone, one can predict the molar refraction of many compounds:

methane, $\text{CH}_4$ :	$R = 2.418 + 4 \times 1.100$ ;
chloroform, $\text{CHCl}_3$ :	$R = 2.418 + 1.100 + 3 \times 5.967$ ;
carbon tetrachloride, $\text{CCl}_4$ :	$R = 2.418 + 4 \times 5.967$ ,

etc. Refractions can be measured with great accuracy and when necessary extremely small differences can be determined.

In view of dispersion anomalies, which, as was explained earlier, occur at frequencies close to the natural frequencies of absorption, refraction should be measured in a region far removed from the absorption bands.

Indexes of refraction are measured by means of refractometers. Most refractometers measure the angle of refraction of a beam of light (emerging from a material under investigation) which impinges on the surface of a prism made of glass with a higher  $n$ .

If a bundle of rays with an angle of incidence of  $0^\circ$  to  $90^\circ$  reaches the boundary between the material under investigation and the glass, the refracted rays will lie between  $0^\circ$  and a certain critical angle  $\alpha$  the sine of which will be equal to the ratio of the index of refraction of the material under investigation to that of the glass of the prism (see Fig. 299). The critical angle is indicated by a sharp line in the focal plane of the tube.

To measure the index of refraction of a liquid, we must place a 0.5-mm layer of it on the surface of the prism. When measuring solids, the material must make close contact with the surface of the prism. Optical contact is

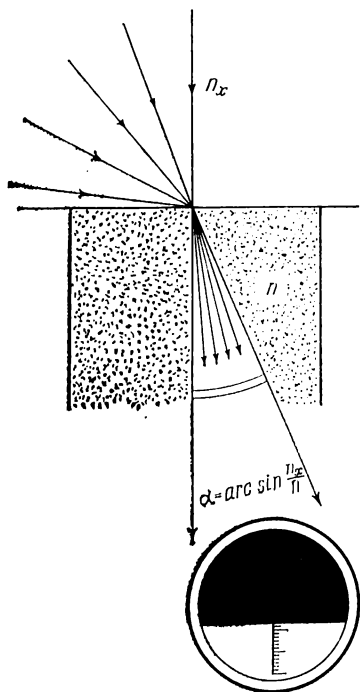


Fig. 299

achieved by using drops of an appropriate liquid between the surface of the

\* These are the contributions of a given atom to the value of a given physical quantity.

prism and that of the material under investigation. The index of refraction of a powder may be determined by immersing the powder in a liquid whose index of refraction is the same as that of the powder.

#### Sec. 274. PYROELECTRIC AND PIEZOELECTRIC MATERIALS

A crystal which does not have a centre of inversion included in its symmetry elements may possess a number of interesting properties. Such crystals may have an electric moment (polarisation vector) in the absence of an external field.

First, let us direct our attention to crystals which become polarised with homogeneous deformation\*. This property is characteristic of piezoelectric crystals, which were discussed in Sec. 45.

The occurrence of polarisation on compression, extension, etc., shows that homogeneous deformation results in the creation of a special, i.e., single (not multiplied by the number of symmetry elements) direction. Such behaviour is not possible when a crystal has a centre of inversion. Homogeneous deformation cannot

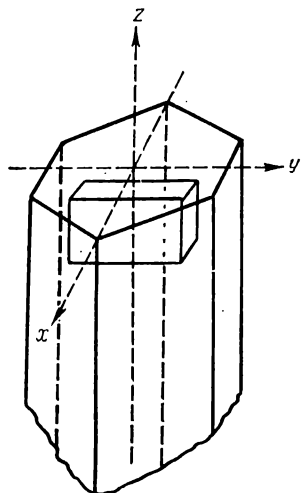


Fig. 300

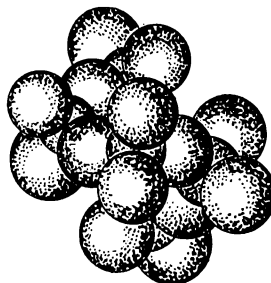


Fig. 301

eliminate a crystal's centre of inversion. At the same time, the existence of a centre of inversion is incompatible with the existence of a special direction such as the polarisation vector direction. Any crystal that does not have a centre of symmetry may possess piezoelectric properties. Nevertheless such properties are not found in many crystals of this type. This may be due to the fact that instruments are not sufficiently sensitive. However, we may conceive of a noncentral symmetric structure in which a homogeneous deformation does not displace the centre of gravity of positive charge relative to the centre of gravity of negative charge. Close examination shows that the piezoelectric effect is not possible in one of the noncentral symmetric groups of symmetry.

The most common piezoelectric material is quartz. Figure 300 shows how piezoelectric plates may be cut from a quartz crystal.

The nature of atom displacements can be assessed from Fig. 301. The structure of quartz may be pictured as a compact packing of oxygen ions in the empty spaces

\* This is a deformation in which all volume elements are deformed in the same manner.

of which silicon atoms are located. The oxygen ions carry a negative charge and the silicon ions a positive charge. A silicon atom is surrounded by four oxygen atoms. Electrification of quartz occurs when it is compressed along the polar axes. Pressure applied along axes lying in the plane of the figure results in the displacement of its positive charge relative to its negative charge. Pressure along an axis of the third order (a nonpolar direction perpendicular to the plane of the figure) is ineffective.

Atom displacements cannot be shown in the figure. These displacements are quite negligible and cannot be detected by objective methods (e.g., X-ray structural analysis). The piezoelectric constant of quartz, i.e., the magnitude of the polarisation vector for unit pressure, is equal to  $6.5 \times 10^{-3}$  CGS unit. The volume of a unit cell of quartz is equal to  $112 \times 10^{-24}$  cm<sup>3</sup>. Since there are three SiO<sub>2</sub> molecules in a cell, the number of molecules in a unit volume is equal to  $2.7 \times 10^{22}$  and, therefore, the dipole moment per molecule for unit pressure is equal to  $2.4 \times 10^{-30}$  CGS unit. The charge of a molecule is equal to  $14 + 2 \times 8 = 30$  electron charges. Therefore, the displacement of the centre of gravity of positive charge relative to the centre of gravity of negative charge is a negligible quantity of the order of  $10^{-18}$  cm, which gives 0.1 Å at a pressure of 1,000 atm.

Piezoelectric crystals include a class of materials known as *pyroelectric* crystals. Such crystals are naturally polarised under normal temperature and pressure. Usually this effect is masked by the free surface charge which accumulates along the boundaries of the crystal, but it may be detected when the temperature of the crystal is raised. Hence the designation pyroelectric (pyro means fire).

Pyroelectric crystals have even more restricted symmetry. Only a crystal having a special axis can be termed pyroelectric. Thus, the mere absence of a centre of symmetry is insufficient. The significance of this condition is evident. The presence of natural polarisation indicates that such a special direction is present in pyroelectric crystals, while in the case of piezoelectric crystals such a direction appears only under the action of mechanical deformation. One of the most common pyroelectric substances is tourmaline.

A pyroelectric crystal has a very strong internal electric field. Therefore, the superposition of an external field does not change the polarisation of such a crystal, i.e., the polarisation cannot be increased, decreased or rotated. Such a crystal is polarised to saturation—all particle dipole moments are parallel. In the theory of ferromagnetism (see Sec. 279, which may be read with profit at this point), a region in which the magnetic moments of atoms are parallel is called a domain. The same term is applied to a region in which the electric dipole moments of all particles are parallel. A pyroelectric crystal usually constitutes a single domain.

#### Sec. 275. FERROELECTRIC CRYSTALS

Of great importance for engineering is the class of ferroelectric crystals possessing the following peculiarity: an increase in temperature leads to the loss of their pyroelectric properties. Maximum permittivity (dielectric constant) is observed at the transition temperature (the so-called Curie point). The temperature relationship of  $\epsilon$  satisfies the law

$$\epsilon = \frac{C}{T - \Theta},$$

where  $\Theta$  is the Curie temperature.

Thus, in ferroelectric crystals (a typical representative of this class of materials is Seignette salt) there takes place a transition from a polarized ordered state to an unpolarized state.



There exists a close similarity between ferroelectric and ferromagnetic substances. The same as ferromagnetics, these interesting dielectric substances have very large dielectric constants (values of several hundreds or even thousands), pronounced hysteresis effects, and Curie points. Along with ferroelectric crystals in whose domains all the dipoles are parallel, there exist antiferroelectric substances with alternated directions of dipoles.

The discussion of Sec. 279 is completely applicable here and will not be repeated. The points made about the effect of a field, polarisation by the displacement of domain boundaries, and the reasons for the division of a crystal into small domains—all apply to ferroelectric materials as well.

Today we know over a hundred of various substances with ferroelectric properties. Various crystals may have somewhat different mechanisms of polarisation.

A large number of ionic pyroelectric crystals are ferroelectric. Several of these are particularly suitable to demonstrate ferroelectric properties. A profound study was given to the phenomena occurring in the family of substances designated by the general formula  $ABO_3$  (so-called perovskites). Barium titanate,  $BaTiO_3$ , belonging to this family, is typical in this respect. At  $120^\circ\text{C}$ , barium titanate loses its special properties and becomes an ordinary dielectric. At temperatures above  $120^\circ\text{C}$ , this substance has the simple unit cell shown in Fig. 302. The cell is cubic; at the centre there is a titanium atom, at the corners of the cube barium atoms, and at the face centres oxygen atoms. The cell has a central symmetric structure; above  $120^\circ\text{C}$ , the crystals no longer exhibit pyroelectric properties. When the temperature is reduced a phase transition occurs and the structure changes: one of the cube edges becomes 1 per cent longer than the other two and the cube is transformed into a tetrahedron. In this process, the titanium atom is displaced in the direction of one of the oxygen atoms. This now becomes the special direction and the polarisation vector will be parallel to this line. It is clear that a barium titanate crystal has three directions of weak polarisation, rather than one, since the displacements along the three axes of the cube are equal.

When the substance is cooled below the Curie point (which is  $120^\circ\text{C}$  for a barium titanate crystal), different regions of the crystal may be transformed into domains with different orientations. A crystal which has acquired a domain structure is in a state of mechanical stress, i.e., some portions of the crystal are compressed and others extended. Strictly speaking, a domain crystal is not a monocrystal, since three-dimensional long-range order throughout the crystal is no longer present.

On further decrease in temperature, barium titanate undergoes yet another phase transformation at about  $+10^\circ\text{C}$ , but retains its ferroelectric properties.

Seignette salt behaves differently. It possesses the ferroelectric properties only within a narrower temperature interval: from  $-20^\circ\text{C}$  to  $+24^\circ\text{C}$ .

Let us consider in more detail the distortions of the symmetric cubic structure which appear when the substance is cooled to a temperature below  $120^\circ\text{C}$ . Neutron-diffraction measurements show that the deformation measured with respect to the lattice of barium ions consists in displacements of titanium ions by  $+0.05 \text{ \AA}$  and oxygen ions by  $-0.10 \text{ \AA}$  and  $-0.05 \text{ \AA}$ . Unequal displacements of oxygen ions just show that one of the cube edges becomes singular due to ferroelectric transformation accordingly as the cubic system turns into a tetragonal system.

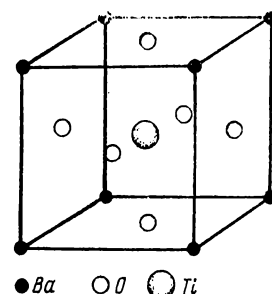


Fig. 302

Between negatively charged oxygen ions and a positive titanium ion a dipole arm is formed. If we project all the charges onto the singular direction, assuming that the atoms of barium and titanium lack two and four electrons, respectively, and the oxygen atoms have two excessive electrons, then an elementary cell will get an electric dipole of  $0.15 \times 10^{-18}e + 0.05 \times 10^{-18}2e = 0.7 \times 10^{-18}e = 3.5 \times 10^{-18}$  CGS unit. Since the volume of the elementary cell is equal to  $64 \text{ \AA}^3$ , the polarisation vector will be equal to  $\frac{1}{64 \times 10^{-24}} \times 3.5 \times 10^{-18} = 5 \times 10^4$  CGS units. A direct measurement of the polarisation vector yields, within the experimental error, the same magnitude.

# Magnetic Substances

## Sec. 276. THREE GROUPS OF MAGNETIC SUBSTANCES

Substances may be classed into three groups in accordance with their magnetic properties: diamagnetic, paramagnetic and ferromagnetic. The values of diamagnetic susceptibility lie in the range of  $-13 \times 10^{-6}$  (bismuth) to  $-0.8 \times 10^{-6}$  (copper). Paramagnetic bodies are characterised by positive susceptibility—for example,  $0.4 \times 10^{-6}$  (potassium) and  $320 \times 10^{-6}$  (iron chloride). Ferromagnetic bodies are characterised by large values of permeability. These are hundreds and even thousands of times greater than those of other bodies. Let us examine the structural features which explain these great differences in magnetic properties for substances which otherwise do not show great differences in properties.

Diamagnetism, it will soon be seen, is a universal property of all bodies inasmuch as they consist of electrons. The above values show that diamagnetic properties are weaker than paramagnetic ones and, *a fortiori*, weaker than ferromagnetic properties. Diamagnetic properties may be detected only in the absence of properties resulting in positive magnetism. Paramagnetic and ferromagnetic bodies have diamagnetic properties, but they are obscured by the stronger positive paramagnetism. Thus, diamagnetism exists for any system containing electrons. On the other hand, positive magnetism arises only in bodies the atoms of which possess a magnetic moment. The phenomenon of paramagnetism is very similar to the process of electrification of a dielectric, which consists of rigid dipoles possessing a constant dipole moment.

The presence of a magnetic moment in atoms is also a necessary condition for the existence of ferromagnetic properties. However, the peculiarities of ferromagnetic substances are due to a very specific property, viz., the formation within a body of vast regions—domains—within which the magnetic moments of thousands of millions of atoms are arranged parallel to one another.

## Sec. 277. DIAMAGNETISM

Diamagnetism is a direct consequence of the tendency for an electron to move in a circle in a magnetic field.

In a magnetic field with an induction  $B$ , an unbound charged particle moves in a circle with an angular frequency  $\omega = \frac{eB}{mc}$ . It can be rigorously proved that the action of a magnetic field on an electron moving in a central field—in particular, in the field of an atomic nucleus—produces an analogous effect: the electron will move in a circle about a line of force, but at one-half the frequency, viz.,  $\frac{eB}{2mc}$ . This motion is superimposed on other motions which may be performed by the electron, the chaotic motion of particles of the electron gas or the motion of the electron about an atomic nucleus.

The fundamental considerations discussed on p. 384 showed that such motion may be equated to a circular electric current. When the magnetic field is switched on, the electrons begin to rotate about the magnetic field and each produces an elementary current

$$I = \frac{ve}{2\pi r} = \frac{e\omega}{2\pi}.$$

Multiplying this value by the area of the circle described by an electron in its motion about a line of force, we obtain the value of the diamagnetic moment created by one electron:

$$M = -\frac{1}{c} \frac{e\omega}{2\pi} S = -\frac{e^2}{4\pi mc^2} SB,$$

The reason for the minus sign is clear from Fig. 303, the direction of the moment is opposite to that of the field.

When a system consists of a large number of electrons, we must take the summation of the above expression with respect to all the electrons:

$$M = -\frac{e^2}{4\pi mc^2} \sum_i S_i B.$$

Since by definition (see p. 218) magnetic susceptibility is equal to the ratio of magnetic moment per unit volume (or unit mass or mole) to induction,

$$\chi = -\frac{Ne^2}{4\pi mc^2} \sum_i S_i.$$

If  $N$  is Avogadro's number,  $\chi$  represents molar diamagnetic susceptibility (in comparing with the results on p. 218, note that  $\chi = \frac{\kappa}{\mu}$ ).

Thus,  $\chi$  is given by the areas circumscribed by electrons in their secondary motion in the magnetic field. In principle, this computation can be made if we know the wave function of the system, i.e., in the final analysis, the electron density. Actually, since the computation is very cumbersome, the diamagnetic susceptibility is determined experimentally.

It should be emphasised that diamagnetic susceptibility is determined by the electron structure of the system and does not depend (at least for atoms and molecules) on external conditions, including temperature.

Diamagnetic susceptibility, like molecular refraction, possesses additivity. If the diamagnetic susceptibility is taken for a mole of substance, the susceptibility  $\chi$  of a molecule may be expressed with considerable accuracy as

$$\chi = \sum n_A \chi_A,$$

where  $n_A$  is the number of atoms of type  $A$  in the molecule and  $\chi_A$  is the increment for the given atom.

For purposes of illustration, we can use the same example as for refraction (see p. 524). C, H and Cl atoms have the increments 7.4, 2.0 and 18.5 ( $\chi_A \times 10^6$ ), respectively. Thus, we obtain 15.4 for methane, 64.9 for chloroform, and 81.4 for carbon tetrachloride. These values are in close agreement with experimental results.

The significance of this additivity consists probably in the following: outer electrons weakly affect diamagnetic susceptibility. In so far as additivity is realised, diamagnetic susceptibility is an atomic rather than a molecular property.

Diamagnetic susceptibility, as indicated in the preceding article, is a property associated with substances the atoms and molecules of which do not have a constant magnetic moment. Such particles include in the first place atoms and ions with completed shells—the ions  $F^-$ ,  $Cl^-$  and  $Na^+$  and atoms of the noble gases. Atoms and ions which in addition to a completed shell contain two more  $s$ -electrons with anti-parallel spins, e.g., Zn, Be, Ca and  $Pb^{++}$ , are also diamagnetic.

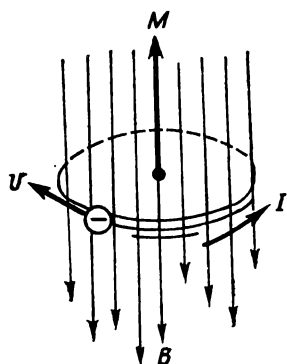


Fig. 303

The group of diamagnetic molecules is incomparably larger than the group of paramagnetic molecules. The latter exists more in the nature of exceptions. This is due to the fact that practically all molecules have valent bonds formed by a pair of electrons with anti-parallel spins. Usually, the total moment about a nucleus, as well as the spin moment, equals zero in such molecules. Thus, bodies consisting of atoms and ions such as those cited above and practically all bodies the building blocks of which are molecules—therefore, practically all organic substances—are diamagnetic.

Diamagnetic susceptibility describes the electron cloud of a molecule. If the distribution of electrons in a molecule is strongly anisotropic, its magnetic susceptibility is also anisotropic. The anisotropy of diamagnetic susceptibility is manifested particularly in molecules of the aromatic compounds. For example, in benzene,  $\chi_{\parallel}$ , the molar diamagnetic susceptibility in a direction lying in the plane of a benzene ring, equals  $-37 \times 10^{-6}$  cm<sup>3</sup>/mole and  $\chi_{\perp}$ , the molar diamagnetic susceptibility in a direction perpendicular to the plane of a ring, equals  $-91 \times 10^{-6}$  cm<sup>3</sup>/mole; in naphthalene  $\chi_{\parallel} = -40 \times 10^{-6}$  cm<sup>3</sup>/mole and  $\chi_{\perp} = -190 \times 10^{-6}$  cm<sup>3</sup>/mole. Anisotropy may be detected by measuring crystals oriented in different directions in the field. Measurements of powders, liquids and gases yield a value of magnetic susceptibility for an averaged orientation.

#### Sec. 278. PARAMAGNETISM

A substance has paramagnetic properties if the atoms, ions or molecules of which it consists possess a magnetic moment. A magnetic moment is due either to the uncompensated spins of electrons in the atomic system or to the motion of electrons about nuclei, or both.

As was explained earlier (see p. 388), a magnetic moment resulting from spin is related to angular momentum as follows:

$$\mu_s = 2\mu_B \sqrt{s(s+1)},$$

and a magnetic moment resulting from the motion of electrons about a nucleus is related to angular momentum as follows:

$$\mu_L = \mu_B \sqrt{L(L+1)}.$$

Here,  $\mu_B$  is the Bohr magneton and  $s$  and  $L$  are, respectively, the total spin momentum and the total angular momentum for motion about a nucleus, taken for an atom or molecule as a whole. As previously,  $s$  and  $L$  are expressed in units of  $\frac{h}{2\pi}$ . When paramagnetism is due to both effects, the formula for the magnetic moment of an atom or molecule takes the form

$$\mu = g\mu_B \sqrt{J(J+1)},$$

where  $J$  is the quantum number of the total quantum momentum, i.e., the vector sum of  $L$  and  $s$ , and  $g$  is the *Lande factor*, which depends on all three quantum numbers. Incidentally, the proximity of  $g$  to 1 or to 2 (established experimentally) is an excellent indicator of the origin of the magnetism of a given substance.

Paramagnetic atoms and ions include particles having one electron over and above a completed shell (e.g., atoms of the alkaline metals), atoms of the transition elements, ions of the rare earth elements with incomplete shells, etc.

Most molecules, as already indicated, are diamagnetic. Molecules of oxygen and sulphur, which are paramagnetic, are exceptions and have a total spin equal

to 1. The magnetic moment obtained experimentally is in close agreement with the value calculated by means of the formula

$$\mu = 2\mu_B \sqrt{2}.$$

The presence of paramagnetism is proof of the fact that the molecules contain unpaired electrons. This circumstance makes the measurement of the magnetic properties of molecules of great interest to the chemist. The so-called free radicals, which are chemical compounds with an unpaired electron, possess paramagnetic properties. Free radicals are created in a number of instances in chemical reactions, and the measurement of magnetic susceptibility is a possible method of studying the course of chemical reactions.

How is the value of the paramagnetic moment of a molecule related to that of magnetic susceptibility? In paramagnetic bodies located outside a magnetic field, the magnetic moments are distributed randomly with respect to direction, and the total magnetic moment of a substance is equal to zero. When a field is switched on, the atoms (or molecules) will tend to rotate in such a way that their magnetic moment coincides with the direction of the field. As a result, equilibrium is established between two tendencies: the ordering action of the field and the tendency to thermal randomness. The reasoning used on p. 521 to derive the value of the polarisability of a substance consisting of rigid electric dipoles is completely applicable here. Therefore, like in that case, the relationship between the magnetic moment of an atom (or molecule) and the paramagnetic susceptibility of an atom is given by the expression

$$\chi_{atom} = \frac{\mu^2}{3kT}.$$

In contradistinction to diamagnetic susceptibility, the paramagnetism of a substance depends on temperature. To be sure, the situation here is somewhat more complex than in the case of dielectrics. This is due to the fact that the electric moment of a molecule is a constant, while the magnetic moment of a molecule (or atom) may vary considerably with the temperature. Paramagnetic moment is related to quantum numbers, and the distribution of molecules according to state may depend greatly on temperature. Therefore, the simple law that magnetic susceptibility is inversely proportional to temperature (the Curie law) may not be valid in the case of paramagnetic substances.

#### Sec. 279. FERROMAGNETISM

**Domain.** A small number of substances possess marked (using coarse observation methods) magnetic properties. These substances include iron, cobalt, nickel, gadolinium, compounds of these elements, and certain compounds of manganese and chromium. Since iron is the most important of these, such substances are said to be ferromagnetic.

Atoms of a ferromagnetic substance have a magnetic moment, which, moreover, is caused by spin (at least, basically). However, it is not this feature that distinguishes it from a paramagnetic substance. The main characteristic of a ferromagnetic substance is its domain structure. A *domain* is a region which is magnetised to saturation, i.e., a region in which all atoms are arranged with their magnetic moments parallel. Since the linear dimensions of domains are usually of the order of 0.01 mm, they may be observed by means of an ordinary microscope.

Domains exist in a ferromagnetic substance in the presence as well as in the

absence of a field. In order to observe domains, we place a drop of colloidal suspension—a finely divided substance such as magnetic ( $\text{Fe}_3\text{O}_4$ )—on the polished surface of a ferromagnetic monocrystal. Colloidal particles become concentrated close to the boundaries of the domains since strong local magnetic fields exist along such boundaries (as in the case of any bar magnet) and they attract the grains of magnetite (see Fig. 304).

First, let us consider certain problems arising in connection with one domain; then we shall study the arrangement of domains in crystals; and finally we shall examine the process of magnetisation of a ferromagnetic substance.

The orientations of the magnetic moments of atoms forming a single domain are not arbitrary. Every crystal of a ferromagnetic substance has a particular crystallographic direction along which it is most easily magnetised. In hexagonal cobalt this is a single direction—the hexagonal axis. In cubic iron this direction is the edge of a cube. This means that there are three directions of easiest magnetisation and accordingly three directions of magnetic moments of domains. In cubic nickel the spatial diagonals of a cube are axes of easiest magnetisation, i.e., there are four possible directions of magnetic moment.

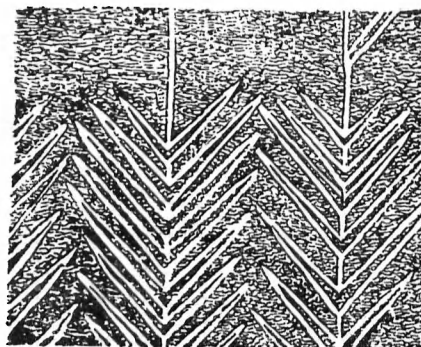


Fig. 304

Why is it that atoms in a ferromagnetic substance arrange themselves so that their magnetic moments are parallel? This is caused by a specific phenomenon—the interchange of positions by electrons. As indicated in connection with a chemical bond, the overlapping of wave functions results in a decrease in energy. Electrons occupy a common space and are able to interchange positions. The tendency of exchange energy to become minimal is the reason for the stability of most chemical compounds. Exchange energy plays an analogous role in the creation of a domain. In the case of a chemical bond, the minimum value of exchange energy is achieved when the spins of interchanging electrons are anti-parallel. However, the general conclusion of quantum mechanics is broader, i.e., the exchange energy may in certain cases be minimal for parallel orientation of spin and in other cases for anti-parallel orientation of spin. In ferromagnetic substances, the spins of atoms contained in a domain have a parallel orientation. Comparatively recently, a new class of compounds—anti-ferromagnetic substances—was discovered. In these substances, stable domain states occur for anti-parallel orientation of spin.

From measured values of the magnetisation of a domain, one can calculate the number of spins per atom involved in ferromagnetism. Such numbers are not whole numbers (for iron 2.2, for cobalt 1.7, for gadolinium 7.1 etc.). It must be concluded that to a certain extent the electrons forming an electron gas are also involved in the creation of ferromagnetism. However, in the main, electrons bound to atoms are responsible for ferromagnetism. In iron, conduction electrons come from the outer  $4s$  shell, while ferromagnetic electrons are in the  $3d$  shell.

The existence of remarkable materials known as ferrites constitutes direct proof of the absence of any connection between conduction properties and ferromagnetism. These materials are semiconductors with a specific resistance of 10 to 11 orders of magnitude greater than iron. Conduction electrons, of course, play no role in the magnetism of these substances. Ferrites are mixed compounds; for

example, manganese ferrite is a 1 : 1 mixture of manganese oxide and iron oxide, and nickel ferrite is an analogous mixture of nickel oxide and iron oxide. Iron oxide contains two iron atoms, and nickel oxide one nickel atom. A crystal of the mixture represents a compact packing of oxygen atoms. The nickel atoms and the two iron atoms fit into the empty spaces. It was indicated on p. 475 that there are two kinds of empty spaces in a compact packing arrangement, viz., tetrahedral and octahedral. An atom which fits into an empty space of the first kind is surrounded by four neighbours, while an atom in an octahedral space has six neighbours. The iron atoms fit into both kinds of spaces. The magnetic moments of the iron atoms are quite ordered, but the moments of iron atoms in tetrahedral spaces point in

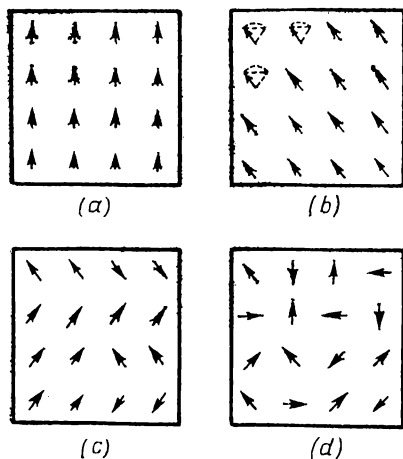


Fig. 305

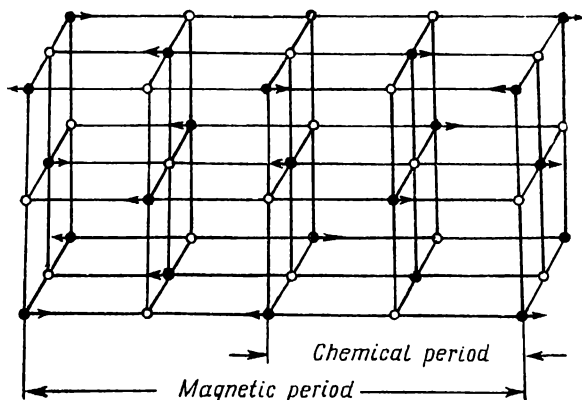


Fig. 306

one direction while the moments of iron atoms in octahedral spaces point oppositely. As a result, the actions of these two systems of moments cancel each other and the magnetic properties of such a mixed oxide result from the magnetism of nickel, the moments of whose atoms are all pointed in one direction.

The presence of exchange energy explains the tendency of atoms to arrange themselves so that their spins are parallel or anti-parallel. Apparently, in ferromagnetic substances, the exchange energy of interaction becomes of prime importance and causes the substance to have a spin arrangement such that the energy assumes a minimum value. In the remaining paramagnetic substances, other components of interaction energy do not allow the exchange energy to make itself felt.

The long-range order of atoms is destroyed at a certain temperature: the crystal becomes fused. Temperature affects the arrangement of the magnetic moments in exactly the same manner. Figure 305 shows schematically how the magnetic moments of the atoms behave when the temperature is raised. At first vibrations are in phase, then disorder begins to prevail, and finally the magnetic order "melts away". Beginning at a definite temperature called the Curie point, in honour of the outstanding French scientist Pierre Curie, the order in the arrangement of arrows disappears and the substance loses its magnetic properties, i.e., the ferromagnetic substance turns into a paramagnetic substance. For iron the Curie point lies at  $770^{\circ}\text{C}$ , for cobalt at  $1,115^{\circ}\text{C}$ , for nickel at  $358^{\circ}\text{C}$  and for gadolinium at  $15^{\circ}\text{C}$ .



In an anti-ferromagnetic substance, the spin of atoms tends to assume an orderly, but anti-parallel, arrangement. The structure of a domain of manganese oxide, which is an anti-ferromagnetic substance, is shown in Fig. 306. Arrows represent the moment of manganese. From the figure, we see that the chemical period of structural repetition is one-half of the magnetic period. At absolute zero each atomic magnet of the anti-ferromagnetic substance is surrounded by atoms with oppositely directed moments. As in the case of a ferromagnetic substance, this order is destroyed at a definite Curie temperature and above this critical point it behaves like a paramagnetic substance.

The existence of various anomalies in the behaviour of a body in passing through the Curie point is indirect evidence of the existence of anti-ferromagnetic properties. Since the Curie point is a point of phase transition of the second kind, a number of properties undergo an abrupt change in passing through it.

Direct evidence of the existence of anti-ferromagnetic properties has been obtained by means of neutron diffraction methods. The scattering of neutrons by a lattice (see Fig. 306) is sensitive to the chemical period, rather than to the magnetic period, of structural repetition.

**Domain Structure of a Crystal.** In examining the domain structure of a ferromagnetic monocrystal by the powder method, which was described earlier, we note that a domain is never very large, i.e., its linear dimensions are usually no greater than 0.01 mm. It is found, moreover, that cubic ferromagnetic substances have extraordinarily symmetric combinations of differently oriented domains. These two circumstances require explanation since, it would seem, that thanks to the ease of magnetisation an entire crystal should be transformed into a single domain.

L. D. Landau and E. M. Lifshits have shown that a domain structure of the kind shown in Fig. 304 is a natural consequence of the existence of different energy forms in a ferromagnetic body. The essence of the theory is illustrated in Fig. 307. The first diagram corresponds to a single domain, the magnetic energy of which is  $\frac{1}{8\pi} \int H^2 d\tau$ . But the energy corresponding to the second configuration is only one-half of this value. In the case of  $N$  parallel domains, the energy will be about  $\frac{1}{N}$  of that for a single domain. However, this dividing process will be advantageous only up to a certain point. Beyond that point, the energy of boundary layers exceeds the decrease in energy associated with the division of a crystal into domains.

The advantage of configurations which consist of domains forming closed circuits is evident. In such cases, a closed magnetic flux circuit is formed and the energy of the field outside the crystal equals zero.

In the case of cobalt, which has a magnetisation direction along its axis, we encounter domains in which the moments are oriented only along the axis of a hexagon. The zero magnetic moment of a body in the absence of an external field

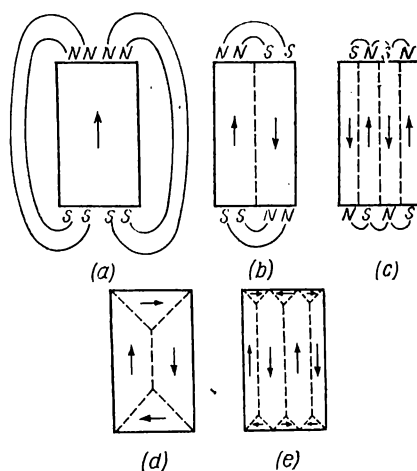


Fig. 307

is realised as follows: half of the domains have one orientation and the other half the opposite orientation.

A few words regarding the boundary between domains. This boundary layer is shown schematically in Fig. 308. We see that in this layer the magnetic moments gradually change direction. The thickness of the layer is determined by the requirement for minimum energy. Two opposite tendencies occur here. On the one hand, it is desirable to extend over a thick layer—this being of greater advantage with respect to exchange energy—the disadvantageous process of spin turning. On the other hand, it is better to complete this process rapidly since in the transition layer the spins are at an angle to the directions of easiest magnetisation.

Now, let us consider what happens in a ferromagnetic substance when an external field is switched on. The magnetisation process may be followed by the powder

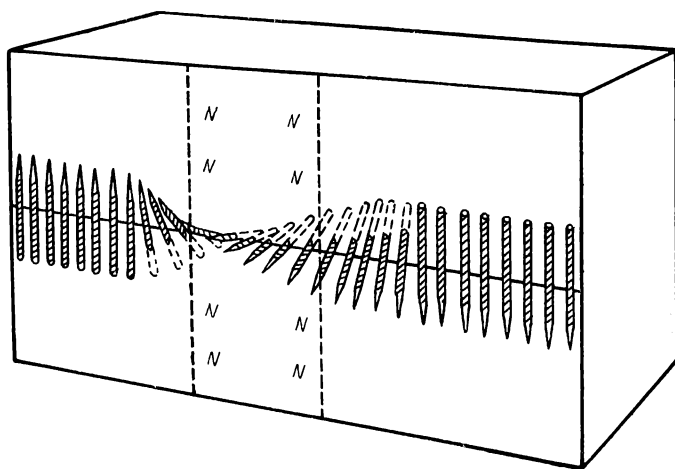


Fig. 308

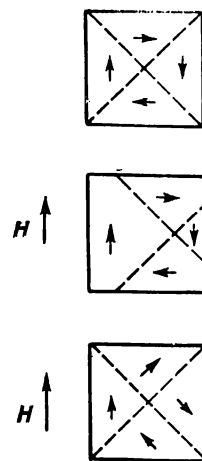


Fig. 309

technique. It transpires that the basic mechanism of magnetisation consists in the growth of a domain, which points in the "required" direction, by means of boundary displacement. The domains which form an acute angle with the field "swallow up" those which form an obtuse angle with the field. At the beginning of magnetisation, domain boundary displacement is reversible, i.e., when the field is switched off the initial boundaries of the domains are restored. Later on, domain boundary displacement becomes irreversible. Finally, when a very high degree of magnetisation is reached, the direction of magnetisation of the domains begins to turn. This is illustrated in Fig. 309.

In polycrystalline substances, the situation is exactly the same (assuming the crystal particles are not extremely small since domains are not formed when the sizes of the particles are less than  $10^{-6}$  cm), i.e., each grain may consist of several domains. In so far as the crystallographic axes of the particles have random orientations in the body, however, the magnetic moments of the domains orient themselves randomly. Thus, the simple magnetisation diagrams which have come down to us from the days of Ampère provide a correct picture of polycrystalline substances.

Hysteresis phenomena inherent to all ferromagnetic materials appear due to the irreversible nature of displacement of domain boundaries during magnetisation.

## Effect of Electron Structure on Properties of Bodies

### Sec. 280. FREE ELECTRONS

Until now, in discussing the structure of solids and liquids, we did not pay particular attention to the role of electrons in the formation of the properties and structure of these bodies. We were able to do this because the electron structure of bodies is by no means always of prime importance. In a number of cases, however, the role of electrons must be taken into account. There are two "kinds" of electrons in a body, viz., bound electrons and unbound (free) electrons. Bound electrons are component parts of a specific atom, ion or molecule. Unbound electrons belong to the entire crystal or liquid and may move quite freely between atoms.

In molecular substances, the picture of electron structure is particularly clear. In most cases, there are no common electrons, i.e., none of the electrons leaves the "bounds" of a molecule. In ionic crystals, the restriction of electrons is not quite so clear. Even according to the classical view of an ionic bond, it cannot be assumed that electron exchange is completely absent. Nevertheless, electrons passing from ion to ion (exchange electrons) in ionic crystals do not behave like free electrons; their displacement in such a crystal consists in the transfer of an electron from one atom to its neighbour. This is quite clear in crystals having a homopolar bond. Diamond is an insulator, although the electrons binding the carbon atoms are by no means restricted to specific positions, but are relayed from atom to atom.

Metals differ quite considerably from all the bodies mentioned above. Here, we encounter electrons for which the term "free" is entirely justified. Electrons are displaced in a metal just like gas particles in a tube filled with obstructions. Atomic products (ions) in a state of thermal vibration act as obstructions. The presence of free electrons is revealed primarily in conduction phenomena as well as all experiments involving the escape of electrons from a body. This type of phenomena could not be explained without considering the peculiar behaviour of common electrons.

It would be incorrect, of course, to assume that the division of electrons into bound and free electrons is absolute. This may rather be considered an idealised division. In solids we may encounter electrons which are bound in various degrees. This became particularly evident when physics assigned a rightful place to semiconductors, which occupy an intermediate position between a system of ideally free electrons and a system of exchange electrons, or a system of electrons bound to molecules. It is now known that any transitional type of structure is possible.

It should be noted that an electron in a solid, just like an atomic electron, obeys the laws of wave mechanics. The representation of an electron as a spherule has validity within the limits imposed by the principle of uncertainty. Usually it is physically meaningless to speak of the path of an electron inside a metal. A description of the electron structure of a body consists primarily in indicating how its electrons are distributed according to energy.

Theory shows that the electrons of a body may be represented as an electron gas, but this statement must be properly qualified. It transpired that it is possible to picture the electrons of a metal as a gas of fictitious particles, the effective mass

of such a particle being dependent on its direction of motion. This point is made in order to caution the reader against making a superficial analogy between an electron gas and a gas consisting of molecules.

#### Sec. 281. ENERGY LEVELS IN A SOLID

The energy levels of a free atom were discussed earlier. Such energy levels may be determined experimentally, i.e., by observing the energy transitions which occur when light is emitted or absorbed. When an atom possesses many electrons, a unique group of four quantum numbers is associated with each electron; according to the Pauli exclusion principle only one electron may exist in a given quantum state. Therefore, energy levels have a limited capacity. The  $s$  levels of an atom may contain two electrons, the  $p$  levels six, etc. This information may be determined experimentally and as a consequence of the fundamental laws of quantum mechanics.

To determine the energy levels of a system consisting of a large number of atoms, we should use both approaches here as well. The fundamental theoretical

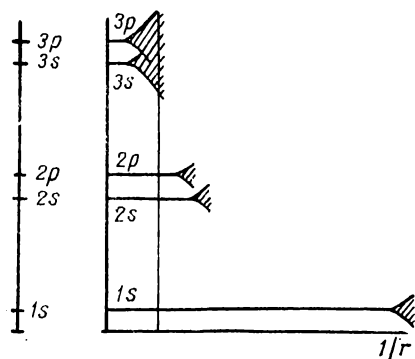


Fig. 310

concepts remain unchanged in the case of a system consisting of thousands of millions of atoms. Therefore, it may be concluded that the number of quantum states in a system consisting of  $n$  atoms will be  $n$  times the number in a free atom. Thus, the Pauli exclusion principle can be satisfied: only one electron will exist in a given quantum state.

An atom never completely loses its identity in a given body. Spectral investigations indicate that significant changes affect only outer, valence electrons, which are responsible for interaction between atoms. Therefore, the quantum states of a solid must be closely related to the quantum states of an atom. Let us consider, for example, the electrons of a  $K$  shell, which is closest to the nucleus of an atom.

It is evident that the state of such an electron can change only very little when atoms are combined in a body. Nevertheless, the Pauli exclusion principle does not allow us to consider all  $K$  electrons as equal. It becomes necessary to assume that  $n$  extremely close  $K$  levels, each of which consists of a pair of electrons of opposite spin, exist in a body consisting of  $n$  atoms.

This reasoning also carries over to the other energy levels. It is assumed that the quantum states of a body are related to that of an atom by the following rule: a body consisting of  $n$  atoms has  $n$  times as many energy levels as an individual atom. Each level of a free atom yields  $n$  close levels in a solid body. This means that the energy levels of a body may be viewed as a system of bands. Each band is a split level of an atom. Therefore, the same designations may be used for the band as in atomic spectroscopy:  $1s$ ,  $2s$ ,  $2p$ , etc. The number of electrons in a band will be, of course,  $n$  times the number of electrons in the corresponding shell of an atom. Thus, in the  $1s$  and  $2s$  bands there will be  $2n$  electrons, in the  $2p$  band  $6n$  electrons, etc.

The width of a band depends on the interaction forces between atoms. This concept is illustrated schematically in Fig. 310. The energy levels of sodium atom are shown at the left and the expansion of the levels into bands in the formation of

a crystal lattice are shown at the right. The quantity  $\frac{1}{r}$  is plotted along the abscissa. Perceptible expansion of the  $1s$  level does not occur since the required interatomic spacings are absolutely unrealisable. The  $2s$  and  $2p$  bands are also practically unexpanded under normal conditions (indicated by a vertical line). On the other hand, the  $3s$  and  $3p$  bands are expanded to such an extent that they overlap. This means that the interaction which occurs between sodium atoms under normal conditions affects only outer electrons. (Sodium has no electrons in the  $3p$  state. Nevertheless, we shall also be concerned with unoccupied energy levels when the exciting energy is sufficient to transfer an electron to such a level.)

What is the significance of the overlapping of the  $3s$  and  $3p$  bands? Actually, our scheme of correspondence between the energy levels of an atom and a solid fails in this case. However, we shall not let this disturb us. The overlapping of the bands signifies that the wave function properties of an electron in the overlapping region differs from the wave function properties of an atomic electron. Thus, the outer electron of a free sodium atom is an  $s$  electron. In liquid and solid sodium the  $3s$  and  $3p$  bands overlap; the behaviour of the outer electrons of sodium differs from the behaviour of an  $s$  electron, i.e., certain special (hybrid) properties appear (such electrons reflect the peculiarities of  $s$  and  $p$  wave functions).

The described behaviour may be established experimentally by means of spectral methods. The presence of an energy band rather than a distinct energy level can be established by examining the transition of electrons from a higher band to a lower one. That which would have produced a sharp line in the case of a free atom now produces a broad spectral band.

It is more convenient to examine transitions from an energy band to a single distinct level—for example, in the case of sodium, transitions to the  $2p$  level. The spectral band obtained in this manner enables us to determine the width of the energy band as well as the electron distribution according to energy. In the case of sodium, this requires that electrons be dislodged from the  $2p$  shell. The frequencies of the resulting transitions lie in the region of soft X-rays (several hundred angstroms) and are very difficult to detect. Special X-ray tubes, in which the anode serves as the material under investigation, are used in such studies.

From measured values of the intensity of the obtained spectral band, we may plot a curve of intensity as a function of the frequency  $\nu$ . But  $\nu = \frac{\mathcal{E}}{h}$  (where  $\mathcal{E}$  is

the transition energy, i.e., the energy relative to the distinct level), and the intensity at a given  $\nu$  is proportional to the number of electrons having an energy  $\mathcal{E}$ . Curves of  $n(\mathcal{E})$  as a function of  $\mathcal{E}$ , where  $n(\mathcal{E})$  is the fraction of electrons in the band having an energy between  $\mathcal{E}$  and  $\mathcal{E} + d\mathcal{E}$ , may be plotted from experimental data. Three typical curves are shown in Fig. 311. The first curve corresponds to an energy band in which the maximum energy is sharply defined. This signifies that all lower energy levels are occupied. The abrupt drop in the curve indicates that the lower levels are filled to capacity (two electrons per level). The second curve is typical of elevated temperatures. In this case, the edge of the band is smeared and

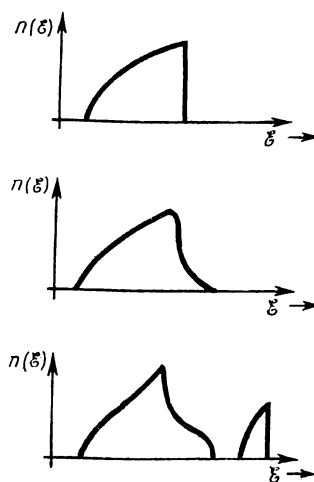


Fig. 311

the order of magnitude of the width of the smeared region is equal to  $kT$ . This means that some of the electrons are in an excited state and can occupy higher levels. The third curve, which shows two nonoverlapping bands, is very interesting. The lower band is filled and the upper one is partially occupied. A forbidden band is located between the two allowed bands.

#### Sec. 282. ELECTRON GAS

It is evident from the preceding article that as far as solid-state theory is concerned only the upper energy bands are of interest, since electrons at lower levels practically do not take part in interactions between atoms. How can the behaviour of upper band electrons be described? Since we are dealing with a very large number of electrons, it is natural to use statistical physics methods and consider an aggregate of such electrons as a kind of gas.

The state of each electron of such a gas may be represented by a point  $(p_x, p_y, p_z)$  in momentum space. The direction of motion of an electron is parallel to its radius vector  $\mathbf{p}$  and the energy of an electron depends on its momentum. In a crystal, the energy of an electron will depend on its direction of motion. Let us disregard this for a moment and assume that the electrons behave like free particles. Despite the fact that this is a rough approximation, i.e., that we neglect the potential energy of the field in which the electrons move and the interaction between electrons, the results give us a good description, at least qualitatively, of the behaviour of the electrons of a solid which form an energy band.

If the electrons are free, the relationship between their energy and momentum is given by the formula  $\mathcal{E} = \frac{1}{2m} p^2$ . This means that in momentum space a surface of equal energy is a sphere. It is customary to call such a sphere a Fermi sphere, after the famous Italian physicist. As indicated in the preceding article  $\mathcal{E}_{\max}$ , the maximum energy of the electrons in a band, may be determined experimentally. We can say, therefore, that the states of an electron gas are contained in a sphere of radius  $p_{\max} = \sqrt{2m\mathcal{E}_{\max}}$ . Thus, it would not be incorrect to call this Fermi surface a surface of maximum energy.

To qualitatively check the validity of this theory, let us estimate from the value of  $\mathcal{E}_{\max}$  the number of electrons in a band. We may reason as follows. According to the principle of uncertainty, the projection of the momentum of a particle in a metal body of linear dimension  $L$  cannot be determined with greater accuracy than  $\frac{h}{L}$ . Therefore, in momentum space, the concept of a point should be replaced by the concept of a cell of volume  $\frac{h^3}{V}$ , where  $V$  is the volume of the metal body under consideration. One of the basic postulates of the theory is that such a cell represents a quantum state and that it can contain no more than two electrons of opposite spins. If there are  $N$  electrons in a volume  $V$  in the band under consideration, then  $\frac{N}{2}$  cells are occupied, i.e., the volume  $\frac{N}{2} \frac{h^3}{V}$ . This is the volume of a Fermi sphere of radius  $p_{\max}$ . Thus,

$$\frac{4}{3} \pi (\sqrt{2m\mathcal{E}_{\max}})^3 = \frac{N}{2} \frac{h^3}{V}.$$

Perfectly reasonable values of  $N$  may be obtained using this equation. This means that the above assumptions are more or less valid.

*Example.* In a metal, the maximum energy, determined experimentally, is  $\mathcal{E}_{\max} \sim 10 \text{ eV} = 16 \times 10^{-12} \text{ erg}$ . Using this value, we obtain  $p_{\max} = \sqrt{2m\mathcal{E}_{\max}} \sim 2 \times 10^{-19} \text{ g cm/sec}$ , i.e., the maximum electron velocity in a metal is

$$v_{\max} = \frac{p_{\max}}{m} \sim \frac{2 \times 10^{-19}}{9 \times 10^{-28}} \sim 2 \times 10^8 \text{ cm/sec}.$$

Hence, the number of electrons in a unit volume is

$$N = 2 \cdot \frac{4}{3} \pi (\sqrt{2m\mathcal{E}_{\max}})^3 \frac{1}{h^3} \sim 10^{23}.$$

The above discussion assumed that the temperature is at absolute zero. At a higher temperature, electrons may pass over into momentum-space cells which correspond to higher energy. Such a transition will take place for electrons located in cells close to a Fermi surface (otherwise too high a transition energy is required, which is unlikely to obtain) and the boundary of the sphere will be broad (not distinct). Only at very high temperatures will the excitation affect low-energy electrons. As the temperature is increased, the degree of degeneracy of the electron gas decreases. An electron gas has a high degree of degeneracy, particularly at low temperatures. The term "degeneracy" signifies that different quantum states have one and the same energy.

The distribution of electrons according to energy at a given temperature may be calculated. This distribution differs from a Boltzmann distribution. According to the Boltzmann law, at absolute zero, the energy of electrons should be equal to zero. From the viewpoint of the new theory, electrons should have a high energy at absolute zero (this follows from the Pauli exclusion principle).\*

On the basis of the Pauli exclusion principle, we can construct a new statistics (Fermi-Dirac statistics), in which the function  $e^{-\frac{\mathcal{E}}{kT}}$  is replaced by the expression

$$\frac{1}{e^{\frac{\mathcal{E} - \mathcal{E}_{\max}}{kT}} + 1},$$

where  $\mathcal{E}_{\max}$  is the maximum possible energy of the electrons at absolute zero. Multiplying this factor by the electron distribution at absolute zero yields the electron distribution at any temperature.

Fig. 312 shows the dependence of the Fermi-Dirac function on  $\mathcal{E}$  when  $kT = 0, 1$  and  $2.5 \text{ eV}$ .

It should be noted that different particles obey different statistics. Molecules obey Boltzmann statistics, photons Bose-Einstein statistics, and electrons (and other particles having a spin of  $\frac{1}{2}$ ) Fermi-Dirac statistics.

\*  $\mathcal{E}_{\max}$  has an order of magnitude of several electron volts, while the average energy of thermal motion ( $kT$ ) is equal to several hundredths of an electron volt. Thus, electrons move rapidly even at absolute zero. The velocity of electrons at absolute zero is 1,000 times as great as the velocity of atoms at room temperature. This should be re-emphasised in order to make it clear that the relationship existing between kinetic energy and temperature in the case of molecules is not applicable in the case of electrons. It follows, moreover, that an electron gas has negligible thermal capacity. The thermal capacity of a body is not affected by the presence of an electron gas.

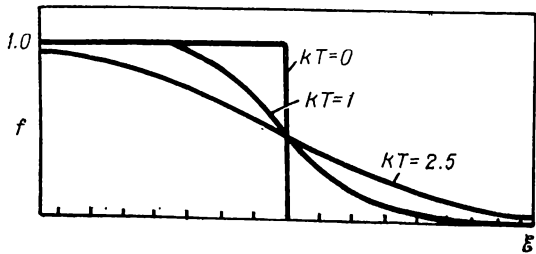


Fig. 312

The difference in statistical approaches consists in the different methods of distribution of particles according to their possible states.

Let us assume that there are two states in which two particles may be possibly located. In Boltzmann statistics, where particles possess individuality, the following possibilities must be considered: (1) both particles in the first state; (2) both particles in the second state; (3) the first particle in the first state and the second in the second state; (4) the first particle in the second state and the second in the first state. Thus, there are four possibilities in all.

In Bose-Einstein statistics, particles are indistinguishable from one another. There are therefore three possibilities: (1) both particles in the first state; (2) both particles in the second state; (3) one particle in the first state and one in the other.

In Fermi-Dirac statistics, the Pauli exclusion principle is taken into account: only one particle may be in a given state. The number of possible distributions is reduced to one, i.e., one particle in each of the two states.

Thus, outer electrons of the atoms of a solid behave like an electron gas. This is a very peculiar kind of gas, i.e., its particles obey Fermi-Dirac statistics.

### Sec. 283. CONDUCTIVITY

In the absence of an electric field, the state of an electron gas is such that the number of electrons moving from right to left is equal to the number moving from left to right. When a field is applied, forces are produced which make the electrons move parallel to the field. The distribution of electrons in momentum space becomes nonsymmetrical with respect to the origin, i.e., a displacement occurs in the direction of the field. An ordered motion which produces an electric current is superimposed on the extremely rapid random motion of the electrons.

For the distribution of electrons to be displaced, electrons must, to be sure, pass to higher energy states. Such a transition is always possible if there are vacancies in the energy band. If the energy band is fully occupied, i.e., if all its levels are occupied by electrons as allowed by the Pauli exclusion principle, the electrons have no place to go—in any case, not until they acquire sufficient energy to make a transition to the next band.

Were it not for the overlapping of bands (discussed above), one could assume that all elements having one valence electron are conductors and all elements giving up two electrons to be shared in the process of forming a solid are insulators. Thus, sodium has one electron at the 3s level. In the formation of a body consisting of  $N$  sodium atoms this level splits into  $N$  levels. At each level there may be two electrons of opposite spin, i.e., a total of  $2N$  electrons. But since we have only  $N$  valence electrons, half of the energy band is unoccupied. Magnesium, the next element in the Mendeleyev periodic table, has two electrons (per atom) at the 3s level. Therefore, in the formation of a magnesium crystal, all levels would be occupied were it not for the overlapping of energy bands.

Investigation of the form of energy bands of various elements shows that the above explanation of the origin of conducting properties is completely valid. Only when the upper band or the merged bands are not fully occupied may the body be classified as a conductor.

The distribution of the electrons of a conducting body in momentum space may be displaced in the direction of a field. Since the number of electrons moving in the direction of a field does not equal the number of electrons moving in the opposite direction, an electric current is produced. In an insulator, all the energy bands are completely occupied. Ordinary field intensities are incapable of creating the forces necessary to transfer electrons to the next higher band, and if we persist



in trying to transfer electrons of an insulator to the next band the dielectric breaks down. The electron distribution maintains its symmetry in momentum space and the number of electrons moving to the left remains equal to the number of electrons moving to the right, i.e., no current flows.

To return to the discussion of conductors, let us now roughly estimate the magnitude of the electrical conductivity of a body which has  $n$  free electrons in a unit volume. By free electrons or conduction electrons, we mean electrons located in unfilled energy bands.

Assume that the motion of an electron under the action of an accelerating force  $eE$  occurs during a small time interval  $\tau = \frac{l}{v}$ . Here,  $v$  is the velocity of an electron and  $l$  the length of its mean free path. The path is traversed at the extremely high random velocity of the electron. The velocity of the ordered motion of the electrons creating the electric current is many orders of magnitude less than the random velocity and, therefore, is not included in the denominator of the expression for  $\tau$ . Motion with an acceleration  $eE/m$  during a time interval increases the electron velocity to  $\frac{eE}{m} \frac{l}{v}$ . Thus, the approximate value of the velocity of the ordered motion of the electrons creating the current is  $u \approx \frac{eEl}{mv}$ .

The density of the electric current is simply the quantity of electricity passing through a unit area per unit time, i.e.,  $j = neu$ . Substituting the above value of  $u$ , we obtain

$$j = \frac{ne^2l}{mv} E.$$

Since Ohm's law in differential form is given by  $j = \sigma E$ , a relation may be obtained for electrical conductivity:

$$\sigma \approx \frac{ne^2l}{mv}.$$

This gives us only a rough estimation of the electrical conductivity. In view of the assumptions made for purposes of simplification, calculated and experimental results will differ by as much as an order of magnitude. However, we are interested merely in a qualitative picture. It can be seen that the conductivity is proportional to the number of free electrons. This number and the length of the free path may vary from substance to substance.

*Example.* If the length of the free path of an electron in a metal is  $l \sim 10 \text{ \AA} = 10^{-7} \text{ cm}$  and  $v$  is of the order of magnitude of  $10^8 \text{ cm/sec}$  (see the example on p. 541), the free path is traversed in a time  $\tau = 10^{-15} \text{ sec}$ .

Assume that the voltage drop along a 1-cm segment of a metal conductor with a 1-cm<sup>2</sup> cross-section is equal to 0.003 V =  $10^{-5}$  CGS unit. Then,  $E = 10^{-5}$  CGS unit and the velocity of the ordered motion of an electron is  $u \sim \frac{eEl}{mv} \sim 5 \times 10^{-3} \text{ cm/sec}$ . The current density is  $j = neu \sim 10^{23} \times 4.8 \times 10^{-10} \times 5 \times 10^{-3} \sim 30 \times 10^{10} \text{ CGS units} = 100 \text{ A/cm}^2$ . This yields quite reasonable values of conductivity:

$$\sigma \sim 25 \times 10^{15} \text{ CGS units} \sim 25 \times 10^{15} \text{ CGS units} \approx 28 \times 10^4 \text{ ohm}^{-1} \text{ cm}^{-1}.$$

If a crystal had an ideal lattice and the temperature approached absolute zero, there would be no restriction on the length of the free path and the material would have no electrical resistance. The electron range is limited by atomic thermal vibrations and the presence of various crystal imperfections. Both factors disturb the ideal periodicity of the field in which an electron moves and result in the scattering of electrons. It follows that the conductivity of a body improves as its tem-

perature decreases and approaches a limit which is determined by the degree of perfection of the crystal lattice.

It can be shown experimentally that the resistance of a metal decreases with temperature. This would indicate that the theory is valid for metals. Moreover, the fact that electrical resistance decreases with temperature is an essential characteristic of metals. The plastic deformation of a metal, the impairment of its lattice by nuclear bombardment, and in general any action serving to damage the lattice will reduce the length of the free path and therefore result in increasing the electrical resistance.

In Part I (p. 167), the thermal conductivity of gases was discussed. It was shown that the thermal conductivity of a gas is proportional to the length of the free path and is given by the formula  $\kappa \sim \rho v l c_p$ . Is this formula useful for the calculation of the thermal conductivity of metals? Electrons are much lighter than atoms and one is justified in assuming that heat is transmitted by electrons which transfer energy from one atom to another. Since the length of the free path is not known, one cannot calculate the coefficient of thermal conductivity. However, it should be noted that the ratio of the coefficient of electrical conductivity to the coefficient of thermal conductivity does not contain unknown parameters and depends only on universal constants and temperature:

$$\frac{\kappa}{\sigma} = \text{const } T$$

(*Wiedemann-Franz formula*). Experimental results agree fairly closely with the value obtained by means of this formula. The following table gives the values of the quantity  $\frac{\kappa}{\sigma T}$  at 0°C for a number of metals.

Metal	Ag	Au	Cu	Mo	Pb	Pt	Sn	Zn
$\frac{\kappa}{\sigma T} \times 10^8 \frac{\text{W} \cdot \text{ohm}}{\text{K}^2}$	2.31	2.35	2.23	2.61	2.47	2.51	2.52	2.31

The theoretical value of this quantity is equal to  $2.45 \times 10^{-8}$ .

#### Sec. 284. SUPERCONDUCTIVITY

Since a crystal always has a considerable number of imperfections, it will usually possess a residual resistance, which is reached at a temperature of several degrees Kelvin, i.e., below this point the resistance does not decrease with temperature. However, there exist about ten metals which behave quite differently. At definite temperatures close to absolute zero, such metals completely lose their electrical resistance. When an electric current is induced in such a superconductor, the current will flow in the circuit for days. This shows that the resistance has not simply decreased, but has dropped abruptly to zero.

Of the pure metals, niobium has the highest temperature (9 K) and hafnium the lowest (0.3 K) at which superconductory properties appear.

It might seem that superconductivity is a property common to all metals, i.e., if the temperature is reduced sufficiently superconducting properties will appear. This is apparently not so. The temperature of numerous materials has been reduced down to 0.03 K without superconducting properties being manifested. The supposition that such properties are not universal is supported by the fact that super-

conducting metals occupy a definite part (the middle) of the Mendeleyev periodic table.

Superconducting materials include, in addition to pure metals, numerous alloys of such nonsuperconducting metals. Moreover, a chemical compound may be a superconductor even though neither of its components is one. Thus, copper sulphide is a superconductor, but copper and sulphur are not. Niobium nitride already reveals superconducting properties at  $30^\circ$  above absolute zero.

The disappearance of electrical resistance at a temperature  $T_h$  is not the only peculiarity of superconductors.

Another mark of a superconductor is its characteristic behaviour in a magnetic field: generally speaking, a magnetic field penetrates such a conductor to a depth of only about  $1,000 \text{ \AA}$ . If very thin films, the behaviour of which is somewhat peculiar, are left out of consideration, we may make the following generalisation: the magnetic field inside a superconductor equals zero.

However, this is true only as long as the applied external field does not exceed a certain critical value  $H_k$ . When this value is exceeded, the superconducting state disappears—the magnetic field penetrates the material and electrical resistance is restored.

$H_k$  is a function of temperature, i.e., it is not constant. At the temperature  $T_h$ , a very weak external field suffices to destroy the superconducting state. Generally speaking at  $T = T_h$  the critical intensity  $H_k$  equals zero.  $H_k$  gradually increases as the temperature is decreased, and at absolute zero it reaches its highest value. For example, the maximum value of the critical field intensity of mercury ( $T_h = 4.2 \text{ K}$ ) is  $412 \text{ Oe}$ .

Electrical resistance is due to the scattering of electrons by the thermal waves of atoms in a crystal lattice. These thermal waves exist, as we know, because of the presence of a zero energy, even at absolute zero. It would seem, therefore, that electrical resistance should not disappear no matter how much the temperature is decreased.

How is it possible then to have thermal scattering of electrons and no resistance to electric current? This problem was not solved until 1937 when it was proved by means of quantum mechanics that electrons in a thin energy layer next to a Fermi surface are able to become "paired" thanks to interaction with the thermal vibrations of a crystal lattice.

It transpired that at low temperatures it is advantageous from the energy viewpoint for two electrons of equal spin magnitude but opposite spin direction to become "united". The words "paired" and "united" have been placed in quotation marks because calculations indicate that the wave functions of these electrons extend over a large distance, viz., of the order of  $10^{-4} \text{ cm}$  (the size of a crystal grain in an ordinary polycrystalline metal). Therefore, the formed pairs should not be viewed as peculiar "molecules", since the bond is implemented over a large distance by means of thermal waves.

It follows from the theory that all electron pairs are identical in the sense that they have the same total momentum. "Matter" consisting of such electron pairs possesses superconduction properties. The formation of electron pairs does not eliminate the thermal scattering of electrons. Superconduction occurs because the scattering of the electrons of a pair ceases to affect the current strength. Thermal scattering can only break up one or another pair or form a new pair from separate electrons, but the magnitude of the current is determined by the total momentum of the electrons, which remains unchanged. Thus, according to this model, the thermal scattering of electrons may produce electric current fluctuations, but it cannot stop the current.

A superconductor contains, in addition to "paired" electrons, an ordinary electron gas, i.e., a gas of individual electrons. Thus, in a superconductor, there exist, so to speak, two fluids—an ordinary fluid and a superconducting one (see p. 498). If the temperature of a superconductor begins to rise from absolute zero, thermal motion will break up more and more pairs of electrons, i.e., the ratio of the ordinary electron gas to the superconducting electron gas will increase. Finally, the critical temperature is reached and the last electron pairs break up.

The new theory provides a quantitative explanation of all of the superconduction phenomena discussed above.

Recent years, the phenomenon of superconductivity has found its application in engineering. Theoretical investigations carried out by Soviet physicists showed that the critical fields in the so-called superconductors of the second kind can reach values up to 3000,000 Gs. The creation of electromagnets with such fields which do not consume energy at all will be an outstanding event for many branches of engineering (an ordinary electromagnet with such a field would require an electric power of about 20,000,000 W—which is just the amount consumed by a town with 20,000 inhabitants). In this case, the winding is made of Nb<sub>3</sub>Sn and also of Nb-Zr and Nb-Ti alloys. Commercial samples of superconducting magnets operating in a helium "bath" produce fields exceeding 100,000 Gs, the field being exclusively homogeneous ( $10^{-6}$  in cm<sup>3</sup>), which is very important for many applications of magnets.

#### Sec. 285. SEMICONDUCTORS

**Properties.** A large group of substances (various elements and chemical compounds), the conductivity of which lies in the broad interval between those of conductors and insulators, are classified under the heading of semiconductors. Since a voltage of 1 V will produce currents of several hundred thousand amperes in a cube of metal having a volume of 1 cm<sup>3</sup>, and currents of the order of  $10^{-10}$  A in insulators under the same conditions, one can see that the interval occupied by semiconductors is extremely large.

The conductivity of substances in this interval has a number of peculiarities which enable us to "recognise" semiconductors.

First, it should be noted that the dependence of conductivity on temperature is opposite to that of metals. The conductivity of semiconductors, in contradistinction to that of metals, may decrease rapidly with temperature. At low temperatures, a semiconductor may become an insulator. The resistance of most semiconductors is considerably more sensitive to changes in temperature than metals. Compact temperature meters of high sensitivity may be constructed using semiconducting thermal resistors (thermistors).

A second important feature of semiconductors is that in a number of cases they may possess positive (*p*) as well as negative (*n*) conductivity. The terms positive and negative are used in the following sense: if the current is due to the motion of positive charges the conductivity is called positive and if it is due to the motion of negative charges it is called negative. Thus, metals have negative conductivity since the current is due to the motion of electrons. Both kinds of conductivity occur in semiconductors. This effect seemed strange at first since the flow of current in a semiconductor is not associated (as in an electrolyte) with the displacement of ions, and the question of the nature of positive carriers of current remained open for a period of time.

The sign of the current carrier may be determined in a number of ways. Let us examine the most convincing evidence, which is based on a study of the forces

exerted by a magnetic field on current-carrying particles (*Hall effect*). If an electric current flows along a plate which is placed perpendicular to magnetic lines of force then, on a charged particle  $e$  moving with a velocity  $u$ , a force  $F$  will be exerted in a direction perpendicular to the field and current (see Fig. 313). In other words, an electric field of intensity  $E = uB$  will be created in such a direction (see p. 224).

A potential difference  $U = uBd$  is produced between the plate faces perpendicular to the created field. The sign of this potential difference is determined by the sign of the charge carrier.

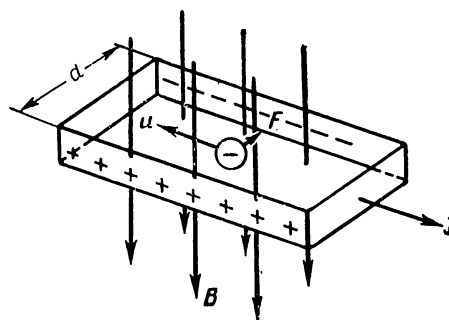


Fig. 313

*Example.* Consider a  $1 \times 2 \times 0.5$  cm<sup>3</sup> semi-conducting plate in a magnetic field  $B = 1,000$  Gs. Assume that the conductivity  $\sigma$  of the plate is equal to  $3 \text{ ohm}^{-1}\text{cm}^{-1}$  (zinc oxide). If a potential difference of 1 V is applied between the plate faces which are separated by a distance of 2 cm,

an electric current of density  $j = \sigma E = 3 \times \frac{1}{2} = 1.5 \text{ A/cm}^2$  will flow through the plate.

Experiments show that a potential difference  $U = 0.12 \text{ mV}$  is produced between the lateral surfaces of the plate. The sign of the Hall effect (see Fig. 313) indicates that the charge carriers are electrons. The velocity of their ordered motion is

$$u = \frac{U}{Bd} = \frac{0.12 \times 10^{-3} \text{ V}}{0.1 \frac{\text{V sec}}{\text{m}^2} \times 10^{-2} \text{ m}} = 0.12 \text{ m/sec} = 12 \text{ cm/sec.}$$

Note that this velocity is more than 1,000 times as great as the velocity of the ordered motion of conduction electrons in a metal (see the example on p. 543). The number of conduction electrons per unit volume of semiconductor is

$$n = \frac{j}{eu} = \frac{1.5 \text{ A/cm}^2}{1.6 \times 10^{-19} \text{ C} \times 12 \text{ cm/sec}} = 8 \times 10^{17} \text{ cm}^{-3}.$$

The low value of  $\sigma$  is due to the fact that this  $n$  is about a millionth of that of a metal.

A final important feature of semiconductors is their extreme sensitivity to impurities, which not only affect their conductivity (an impurity of the order of one per cent may change the conductivity of a semiconductor by a factor of a million or more), but may change  $n$ -conductivity into  $p$ -conductivity and vice versa.

The most important semiconductors already of practical significance include germanium, silicon, selenium, antimony alloys (containing indium, cadmium and zinc), and copper and titanium oxides.

**Interpretation of Properties.** Most characteristics of semiconductors can be easily explained by means of an energy level diagram. Insulators have a filled energy band. The next unfilled band is separated from the filled band by a large energy gap. Imagine that the system of levels of a substance is such that the gap between these bands decreases and the energy of thermal motion suffices to transfer electrons from the filled band to the unfilled band. Such a substance will act as a natural semiconductor.

At a given temperature, the number of electrons in the upper band will be determined by the dynamic equilibrium conditions established between bands. Electrons continuously pass from the lower band to an excited state and vice versa. Just as in the case of a saturated vapour, equilibrium will prevail when the number

of electrons moving "upwards" equals the number of electrons moving "downwards".

Again as in the case of a saturated vapour, when the temperature is raised the equilibrium is displaced in the direction of the upper level, i.e., the instantaneous concentration of electrons in the upper band increases. The concentration of free electrons rises sharply with the increase in gap between bands. The probability of surmounting an energy barrier of width  $\Delta\mathcal{E}$  invariably appears as an exponential factor. The approximate concentration of electrons in the upper band at a temperature  $T$  may be determined from the formula  $n \approx 10^{19} \times e^{\frac{-\Delta\mathcal{E}}{2kT}}$ .

If a body has a gap  $\Delta\mathcal{E}$  which is significantly greater than  $kT$ , it belongs under the heading of insulators. For this purpose, it is sufficient for  $\mathcal{E}$  to be 100-200 times

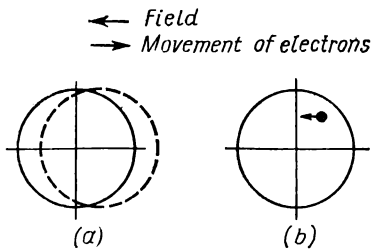


Fig. 314

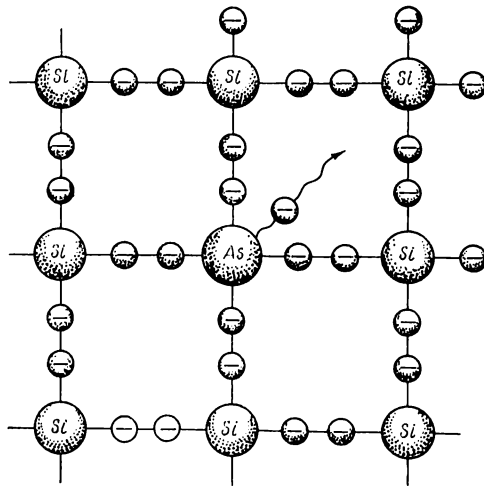


Fig. 315

as great as  $kT$ . At room temperature,  $kT \approx \frac{1}{40}$  eV. When  $\Delta\mathcal{E}$  becomes less than 1 eV, i.e., about 40 times as great as  $kT$ , the number of electrons in the upper band will be sufficient to create measurable currents. When  $\Delta\mathcal{E}$  is of the order of several tenths of an electron-volt, a semiconductor possesses considerable conductivity.

Reference to the conductivity formula  $\sigma = \frac{ne^2l}{mv}$  shows that when the temperature of a semiconductor changes the two factors on which  $\sigma$  depends also change. In the first place, as the temperature increases, the number of free electrons  $n$  increases, but, as previously, the free path  $l$  decreases with increasing temperature. However, experiments show that usually the former effect overshadows the latter.

Until now, we have discussed the conduction properties of the upper band, but have disregarded the lower band, which should also acquire conduction properties since vacancies are formed in it as the result of the transition of electrons to the upper band. This conductivity may be very peculiar in nature.

The creation of conductivity in the upper, partially filled band may be interpreted as a displacement in the distribution of electrons in momentum space in the direction of the field (to the right in Fig. 314a). However, this is not the only way in which ordered motion of electrons may occur. Imagine that the overall shape of the distribution of electrons does not change (see Fig. 314b). However, now

one electron, and now another, close to the Fermi surface is snatched away and a "hole" is formed in momentum space. Under the action of the field, such a hole is immediately filled by a neighbouring electron moving from left to right (in the same direction as in the other diagram). The hole is displaced from right to left. Now, another point representing an electron in momentum space occupies this position and in this manner the hole moves in the opposite direction to that of the electrons. Since holes are formed continually, a continuous positive "hole" current flows.

Thus, in a natural semiconductor, electric current may be viewed as the result of the motion of "holes" in the occupied band as well as of electrons in the upper band. However, the major role in such cases is played by the motion of electrons in the conduction band.

This natural conduction of semiconductors is encountered considerably less frequently than another phenomenon, namely, semiconduction properties under the action of a small percentage of impurities. The role of foreign atoms or other lattice imperfections consists in their contribution to the system of energy levels. Frequently, imperfections create their own level—a narrow energy band between the filled and unfilled bands.

Let us assume that the foreign atoms contribute "surplus" electrons which occupy a narrow band between the main bands. When the temperature increases, electrons pass from the level of the impurities to the conduction band in greater and greater numbers, i.e., conduction increases. Such a semiconductor yields *n*-conduction. A point may be reached (for a low percentage of impurities) at which all surplus electrons are given up. A further increase in temperature will not result in an increase in conduction and from then on the body will act like a metal. Such behaviour may be detected when pentavalent arsenic or phosphorus atoms are introduced into a lattice of quadrivalent silicon or germanium. Figure 315 shows a simplified diagram of the crystal lattice of silicon. If a silicon atom is replaced by an arsenic atom, a "surplus" electron is obtained. This will be a conduction electron.

It is interesting that impurities may result in *p*-conduction. This occurs when an impurity atom has acceptor properties, i.e., can attract electrons. Electrons pass from the filled band to the intermediate level of the impurities and as a result hole conduction occurs in the filled band. Such conduction occurs in silicon containing a trivalent aluminium impurity. If at a number of lattice sites silicon is replaced by aluminium "electron-deficient sites" will exist in the crystal. When a field is applied, an aluminium atom may attract an electron from a neighbouring silicon atom; the electron falls under the action of the electric field and a "hole" moves in the opposite direction.

It should be noted that such "unsophisticated" models of the displacement of an electron are greatly oversimplified in view of the fact that the motion of electrons in a solid satisfies the laws of quantum mechanics.

By adding one or another impurity, we are able to vary the conductivity of materials within very broad limits. We may change *p*-type conductivity into *n*-type conductivity and may significantly change the nature of the temperature dependence of the conductivity.

## Sec. 286. EMISSION OF ELECTRONS

**Work Function of an Electron.** Electrons in a conduction band behave like an electron gas. The surface of a solid acts as the "walls" of the vessel in which this gas is located. To leave the bounds of this surface, an electron must surmount a

potential barrier the height of which is designated by  $\mathcal{E}$ . At absolute zero, electrons have a limiting energy  $W$ . In the model of an electron gas,  $W$  corresponds to a Fermi surface. This is the energy of electrons which at absolute zero are located at the highest level. Thus, in order for an electron to surmount the potential barrier, it is not necessary to impart to it an energy  $\mathcal{E}$ ; it is sufficient to give it an additional energy

$$A = \mathcal{E} - W.$$

The quantity  $A$  is called the *work function* and  $\frac{A}{e} = \varphi$  the *stopping potential*, i.e., the work function expressed in volts (see Fig. 316).

An electron's escape from a metal is impeded by the forces of attraction exerted by positive ions as well as the forces of attraction between the electron and its electrical image. The latter force is equal to  $\frac{e^2}{4x^2}$ , where  $x$  is the distance of the electron from the surface. This force is capable of holding an electron at a considerable distance from the surface and thus forming a layer or electron cloud close to the surface of the body.

If the metal is located in an electric field, the work function decreases by an amount  $e\sqrt{eE}$ , where  $e$  is the electron charge and  $E$  is the field intensity. The intensity of the ex-

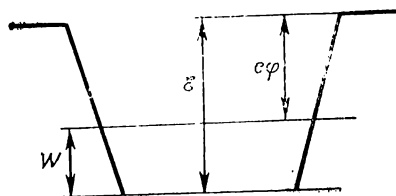


Fig. 316

ternal field must equal  $0.2 \times 10^7$  V/cm if the stopping potential is to decrease by 1 V. (Thus, in ordinary electronic instruments the external field has little effect on the work function.)

The work function is very sensitive to changes in surface properties. It transpired that it is possible to deposit electrically positive atoms or ions (metals, oxygen) on the surface of a cathode. In this manner, a layer of positive charge may form on such a surface. Thus, as may be seen from the following example, the work function can be considerably reduced.

The filament of the heater in electron tubes is usually made of tungsten (work function  $A = 4.9$  eV). By coating the surface of the tungsten with a layer of an oxide of an alkaline earth metal (Ca, Ba, Sr), we may decrease the work function to 1.5-2 eV. This enables us to obtain considerably greater emission at lower filament temperatures.

**Thermionic Emission.** The escape of electrons from a metal upon heating is known as thermionic emission. This phenomenon is basic to the action of heater-cathode tubes. When the temperature is raised, electrons are excited, some of them acquiring a sufficient velocity in the direction perpendicular to the surface of the material to surmount the potential barrier  $\mathcal{E}$ .

An electron gas obeys Fermi-Dirac statistics, according to which the number of electrons having an energy  $\mathcal{E}$  is proportional to the expression  $\frac{1}{e^{\frac{\mathcal{E}-W}{kT}} + 1}$ . But we

are interested in the energies  $\mathcal{E}$  which are considerably greater than the zero-level energy  $W$ . Therefore, it is accurate enough to reduce the above expression to

$$e^{-\frac{W-\mathcal{E}}{kT}} = e^{\frac{-e\varphi}{kT}}.$$

Thus, we may determine the number of electrons having an energy equal to the height of the potential well. It may be rigorously proved that the thermionic emis-



sion current is proportional to this expression. We see from the formula that thermionic current increases extremely rapidly with temperature.

The circuit shown in Fig. 317 may be used to measure the thermionic current. By increasing the voltage, we quickly reach saturation current. (This is the thermionic current referred to above.) The initial portion of the current-voltage curve is space-charge limited (see above). The voltage draws electrons to the plate from the electron cloud and the cathode emits enough electrons to maintain the cloud in equilibrium. In the absence of an external voltage, this equilibrium is determined by the electron emission of the cathode and the negative potential of the cloud. As the voltage is increased, the electron cloud begins to dissipate and emission increases until the voltage draws all electrons of the cloud to the plate. At this point current saturation sets in.

Rigorous analysis yields the following relationship between electronic current and temperature:

$$j = AT^2 e^{-\frac{e\phi}{kT}} \quad (\text{Richardson formula}).$$

For tungsten,  $A = 75 \text{ A/cm}^2\text{K}^2$  and  $e\phi = 4.5 \text{ eV}$ . Let us compare the current density of thermionic emission from tungsten at 500 K and 2,000 K.

At 500 K,

$$j = 75 \times 25 \times 10^4 \times e^{-\frac{4.5 \times 1.6 \times 10^{-12}}{1.38 \times 10^{-16} \times 500}} \sim 10^{-36} \text{ A/cm}^2.$$

In other words, to obtain measurable currents, cathodes of impracticable dimensions (greater than that of the entire land mass of the globe) would be required.

At 2,000 K,

$$j = 75 \times 4 \times 10^6 \times e^{-\frac{5.21 \times 10^4}{2,000}} \sim 1.6 \text{ mA/cm}^2.$$

Such a current is easily measured, but the area of the emitting surface is still too large for most practical purposes.

The picture changes when tungsten is coated with caesium. Now,  $A = 3.2 \text{ A/cm}^2\text{K}^2$  but  $e\phi = 1.36 \text{ eV}$ . At  $T = 2,000 \text{ K}$ ,

$$j = 3.2 \times 4 \times 10^6 \times e^{-\frac{1.57 \times 10^4}{2,000}} \approx 4.8 \times 10^3 \text{ A/cm}^2.$$

Clearly, such current densities would destroy the cathode. Therefore, the required values of current density,  $j \leq 1 \text{ A/cm}^2$ , are attained at lower temperatures ( $\sim 1,300 \text{ K}$ ).

**Secondary Emission.** This refers to emission due to the dislodgement of electrons from a metal under the action of other electrons. Secondary electrons may emerge, taking the same direction as the primary electrons. This shows that primary electrons interact with bound electrons; otherwise, the law of conservation of momentum would be violated. Secondary emission begins at a primary electron energy of the order of 10 eV. Most secondary electrons have an energy of several electronvolts—their energy distribution is practically independent of the energy of the primary electrons.

Primary electrons produce secondary electrons and, in addition, are elastically scattered. The remarkable phenomenon of one primary electron producing several secondary electrons has wide application (e.g., Kubetsky tube).

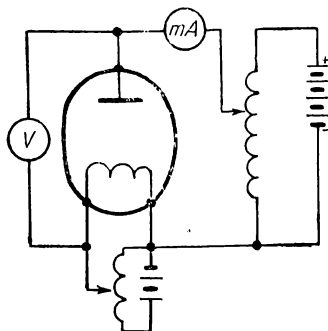


Fig. 317

## Sec. 287. PHOTOELECTRIC EFFECT

**Extrinsic Photoeffect.** This phenomenon may be studied by making the cathode of a vacuum tube out of the material under investigation. Light falling on the cathode dislodges electrons from the material. Electrons reaching the anode generate a photoelectric current, the magnitude of which may be studied as a function of external conditions.

It is the total current taken from the cathode, of course, that is characteristic of the substance. Here too, therefore, we usually operate under conditions of saturation current. If no voltage is applied across the photocell, a weak current generated by electrons leaving the cathode in the direction of the anode flows through the instrument. A weak accelerating voltage is not sufficient to attract all the electrons, but at a certain voltage all the electrons reach the anode and saturation current is obtained.

Experiments show that the photocurrent is strictly proportional to the intensity of the incident light. This is true of light of any frequency which produces a photoeffect.

Moreover, the number of dislodged electrons is exactly equal to the number of photons. One photon may dislodge only one electron. It is not possible for a photon to dislodge several electrons from a substance by a series of energy losses. This important postulate is somewhat difficult to prove in a study of the extrinsic photoeffect, since the extrinsic photoeffect may be accompanied by an intrinsic photoeffect (see below) in which some of the electrons do not leave the bounds of the substance.

The law of conservation of energy and the law of conservation of momentum are obeyed in the interaction of a photon and an electron. The law of conservation of energy (the Einstein equation) takes the form

$$h\nu = \frac{mv^2}{2} + e\varphi$$

where  $\varphi$  is the stopping potential of the electron for the metal (the same as in the thermionic emission experiments). In accordance with the law of conservation of momentum, it may be assumed that the lattice of the metal takes part in the photoelectron interaction process (otherwise electrons could move only in the same direction as the photons).

A photon may produce a photoeffect if its energy is not less than the work function. It follows that each material has a photoeffect limit. The limiting frequency  $\nu_0$  is equal to  $\frac{e}{h} \varphi$  and the limiting wavelength  $\lambda_0$  (in millimicrons)—the retarding photoelectric threshold—is  $\frac{hc}{e\varphi} = \frac{1,236}{\varphi}$ , where  $\varphi$  is expressed in volts. No photoeffect will take place when a substance is irradiated with light of long wavelength. The photoelectric threshold in the case of Li is  $\lambda_0 = 560 \text{ nm}$  ( $5,600 \text{ \AA}$ ), i.e., it is in the yellow region of the visible spectrum; in the case of Cu,  $\lambda_0 = 300 \text{ nm}$  ( $3,000 \text{ \AA}$ ), i.e., the ultraviolet region; and in the case of Al,  $\lambda_0 = 410 \text{ nm}$  ( $4,100 \text{ \AA}$ ), i.e., the violet region of the visible spectrum.

If the energy of a photon is greater than the work function, the surplus goes into the kinetic energy of the electron. Thus, hard radiation can produce very fast photoelectrons.

In order to measure exactly the threshold frequency and the work function, we use a retarding potential method. A small retarding voltage is applied to the photocell (plus terminal is connected to the photocathode) and this voltage is increased until

the current cuts off. This point is reached when  $eV_b = \frac{mv^2}{2}$ . In this manner, we may determine experimentally the dependence of  $V_b$  on the frequency of light:

$$V_b = \frac{h\nu}{e} - \varphi.$$

The plotted curve is a straight line the slope of which is equal to the universal constant  $\frac{h}{e}$ . The threshold frequency  $\nu_0$  and the stopping potential  $\varphi$  are obtained as intercepts along the abscissa and ordinate, respectively (see Fig. 318).

*Example.* If soft X-rays of wavelength  $\lambda = 100 \text{ \AA}$  fall on a copper plate ( $\varphi = 4.1 \text{ V}$ ), the retarding voltage cuts off the photocurrent when

$$eV_b = h\nu - e\varphi = 6.6 \times 10^{-27} \frac{3 \times 10^{10}}{10^{-6}} \times \frac{1}{1.6 \times 10^{-12}} - 4.1 \text{ eV} = 120 \text{ eV}.$$

Therefore, the retarding voltage will equal 120 V.

Another important characteristic of a photocathode material is the spectral dependence of the photocathode. Here, no simple relationship may be applied. The curve begins at the threshold frequency and in many cases increases rather uniformly; one can say that the coefficient of utilisation of photons increases with photon energy. However, in other cases, the spectral curves have well defined maxima which lie within a rather narrow spectral band.

Photocells which utilise the extrinsic photoeffect have broad application. They are used in photoelectric relays, television and cinema sound tracks. Silver, caesium and potassium may serve as photocathodes; antimonycaesium cathodes are widely employed.

In various photo relay applications, the photocurrent need not be proportional to the intensity of light, but the sensitivity of the photocell should be high. In such cases, gas photocells instead of vacuum photocells may be used. This increases the sensitivity tens of times.

**Intrinsic Photoeffect.** When the action of a photon results in the displacement of an electron from a filled band to a level of an impurity or to a conduction level, it is referred to as the intrinsic photoeffect. Under the action of light, this phenomenon may produce conduction electrons and holes in a body. Such conduction electrons and holes will occur in pairs. Strictly speaking there will be a pair of charges for each photon. The phenomenon is extremely complicated by secondary processes which occur within a body as a consequence of the recombination of electrons and holes.

It is clear, therefore, that the intrinsic photoeffect is a phenomenon which is particularly characteristic of semiconductors, but which may also occur in insulators.

Semiconductors which possess this effect are included in current circuits as photoresistors. In the dark, such a body has very low (dark) conduction. Its conduction increases when subjected to light. Energies of several tenths of an electron-volt may be sufficient to produce intrinsic electron transitions. Therefore, the threshold of the intrinsic photoeffect may lie in the far infrared region.

Photoresistors are widely used in signal systems and in automation when it is necessary to amplify or detect very small changes in light intensity.

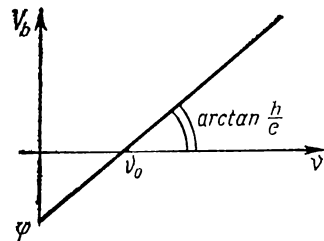


Fig. 318

## Sec. 288. BARRIER LAYERS

In a number of cases, the junction between a metal and a semiconductor or between two semiconductors may have a rectifying action. Even though the boundary between two bodies in close contact (welded or fused) is very narrow, nevertheless it is of finite width; hence the designation *barrier layer*. Such a layer may form at the junction between copper and cuprous oxide or at the junction between selenium and cadmium selenide. Investigations indicate that a barrier layer between two semiconductors is formed when one of the semiconductors is a  $p$ -type conductor and the other an  $n$ -type. Such barrier layers are called  $p$ - $n$  junctions.

Fig. 319 illustrates the rectification provided by barrier layers. The figure shows a typical current-voltage curve. The right branch of the curve is the characteristic for the forward current and the left branch for the reverse current. The forward

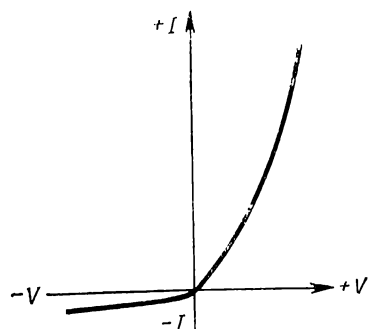


Fig. 319

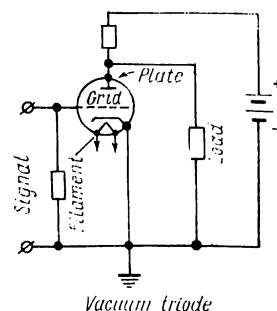
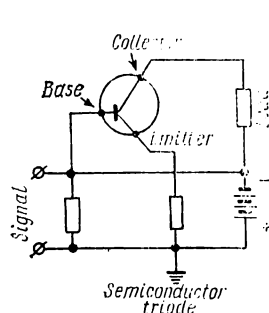


Fig. 320

current increases rapidly with increasing voltage, but the reverse current remains almost constant and has a very low value.

Which is the direction of forward current? Investigations indicate that at a  $p$ - $n$  junction the forward current flows from the  $p$ -type semiconductor, through the junction, to the  $n$ -type semiconductor. This means that holes move toward electrons and electrons move toward the higher concentration of holes. Reverse current flows when the holes and electrons move away from the junction.

In the case of a metal-semiconductor junction, the situation may be described as follows. If the metal forms a junction with a  $p$ -type semiconductor (copper-cuprous oxide), the forward current will flow from the cuprous oxide to the copper. This is to be expected: there are no free electrons in a semiconductor, but in a metal there are an excess of such electrons; therefore, electrons move from the metal to the semiconductor.

The characteristics of barrier layers find wide application in industrial rectifiers. Copper-oxide rectifiers (copper-cuprous oxide) and selenium rectifiers have been produced for a long time. During recent years, tiny germanium and silicon rectifiers—crystal diodes—have been widely introduced. The introduction of impurities into germanium or silicon may transform them into  $p$ -type or  $n$ -type conductors.

A crystal diode consists of a very small germanium (or silicon) crystal, one part of which contains an acceptor-type impurity and the other a donor-type impurity.

Also of interest are crystal triodes, representing a semiconductor system of the  $p$ - $n$ - $p$  or  $n$ - $p$ - $n$  type. If a wire is soldered to each of the three sections of such a tiny triode (the dimensions of crystal "tubes" are of the order of a centimetre), the system may be connected in a circuit just as an ordinary triode tube. A voltage is

applied across the two outer ends, one end serving as the anode and the other as the cathode. The third tap serves as the grid. Such a system of semiconductors has two barrier layers connected in opposition. This is why it behaves like a triode tube. The analogous operation of crystal triodes and ordinary triodes is illustrated in Fig. 320.

Another important use of barrier layers is in the manufacture of photocells, which operate without a voltage source. Coating a semiconductor with a thin layer of metal will produce a barrier layer. A layer of metal may be made so thin that light easily passes through it. When light passes through the metal, an intrinsic photoeffect is produced in the semiconductor. The presence of a barrier layer causes the liberated electrons to move in a definite direction. An electric current will flow when the circuit is closed.

Copper oxide and selenium photocells are manufactured on the basis of this principle. Sulphur-thallium photocells, having a shortcircuit current of the order of 10,000 microamperes per lumen, are at present being used in the Soviet Union. They have an efficiency of transformation of light energy into electrical energy of the order of 1 per cent. Here too silicon and germanium  $p$ - $n$  barrier layers are of fundamental significance. They enable us to manufacture photocells with an efficiency of the order of 10 per cent. This new discovery has placed the utilisation of solar energy on a practical basis.

#### Sec. 289. CONTACT POTENTIAL

When two metals or semiconductors are in contact, there arises a difference of potentials between them. This difference is known as contact potential. To measure this difference, we must remove inclusions, oxide films, etc., and make close contact between the surfaces of two such bodies by soldering, welding or pregrinding. In this manner, we may form a circuit that is broken in one place. Since all points of the body are at the same potential, the contact potential may be determined by measuring the field in the gap. The order of magnitude of the contact potential between two bodies is equal to several tenths of a volt.

The existence of contact potential is easily explained. It is due to the difference in work functions of the two adjacent bodies. It should be recalled that the work function  $A$  is equal to the difference between the energy  $W$  of an electron at the highest level inside a metal (at absolute zero) and the energy of an electron escaping from the metal with zero kinetic energy.

The energy distribution of electrons in the case of two solids placed in contact can be represented as shown in Fig. 321. The upper level is the same for all bodies. The electrons of the body having a lower work function are at a higher energy level. Thus, the conditions are present for the transition of electrons from the first body to the second. This transition is accompanied by the formation of positive charge on the first body and negative charge on the second. At the point of contact, there arises an electric field which impedes the transition of electrons. Finally, equilibrium will be established at a particular potential difference characteristic of the given pair of metals.

This picture depends little on temperature. As the temperature is raised, the energy distribution boundary of the electrons is no longer so sharply defined. Electrons appear at higher levels, but the conditions for the transition of electrons

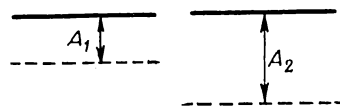


Fig. 321

remain basically the same thanks to a lack of close dependence of the energy of the electrons on temperature.

It is evident from the above description of this phenomenon that any group of solids may be arranged in a definite sequence such that each member of the sequence becomes positively charged with respect to the next. Such a series was first obtained by Volta, the discoverer of contact potential. From the above explanation of this phenomenon, it is clear that a Volta series corresponds to rising work functions, i.e., the motion of electrons between two bodies is in the direction of the one having a higher work function.

Since the contact potential  $\varphi_{12}$  between two bodies may be expressed in the form of a difference of work functions,

$$\varphi_{12} = \frac{1}{e} (A_1 - A_2),$$

it is evident that the potential difference between two bodies may be expressed as the difference of the contact potentials between each of these bodies and a third:

$$\varphi_{12} = \frac{1}{e} (A_1 - A_2), \quad \varphi_{13} = \frac{1}{e} (A_1 - A_3);$$

$$\varphi_{23} = \frac{1}{e} (A_2 - A_3) = \varphi_{21} - \varphi_{31}.$$

Furthermore, it is evident that in a closed circuit consisting of any number of bodies the total contact potential is equal to zero:  $\varphi_{12} + \varphi_{23} + \varphi_{31} = 0$ . No current will flow in such a circuit.

**Electromotive Series**  
(normal potential in an electrolyte solution, volts)

Li	Ca	Na	Al	Zn	Fe	Ni	Pb
-3.01	-2.84	-2.71	-1.66	-0.76	-0.44	-0.23	-0.12
		Cu	Hg	Ag	Pt		
		+0.34	+0.70	+0.80	+1.2		

#### Sec. 290. ELECTROLUMINESCENCE OF SEMICONDUCTORS

Electrons and holes of a semiconductor can combine with photon emission. We can imagine four variants of such processes: transition of an electron of the conduction band to a hole of the valence band and to a hole of the acceptor level, transition of an electron of the donor level to a hole of the valence band and to a hole of the acceptor level.

For a semiconductor to be a light emitter, it is necessary that its structure favours a rapid recombination of electrons and holes and also enables the electrons to be brought to excited states. Such states will be obtained if we succeed in injecting electrons into a semiconductor with a greater amount of holes, i.e., into a *p*-crystal. The same effect is obtained by introducing holes into an *n*-type semiconductor. Finally, we may also resort to injection of holes and electrons into the insulator.

If, by passing an electric current through a semiconductor, we realise one of the above processes, then there will occur direct transformation of current energy into light, i.e., electroluminescence.

*P-n* diodes made of binary semiconductors (such as gallium phosphide or arsenide) turned out to be most convenient for realisation of electroluminescence. An energy level diagram of a diode is shown in Fig. 322*a*. According to the above explanation,

between the  $p$ - and  $n$ -regions of the diode there will set in a potential difference balancing the diffusion of the electrons (indicated by small black circles) into the  $p$ -region and that of the holes (white circles) into the  $n$ -region (Fig. 322a).

When a field is superimposed (Fig. 322b), the barrier becomes lower, the electrons start moving to the right, and the holes in the opposite direction. Favourable conditions for recombinations of all four types are created in the boundary layer. The energy of the obtained photons is, roughly speaking, equal to the gap between the bands.

Of course, the recombination process must not necessarily be accompanied by emission. The corresponding energy can convert into heat as well. If we succeeded in realising an ideal case, then the emission output would exceed the electric energy

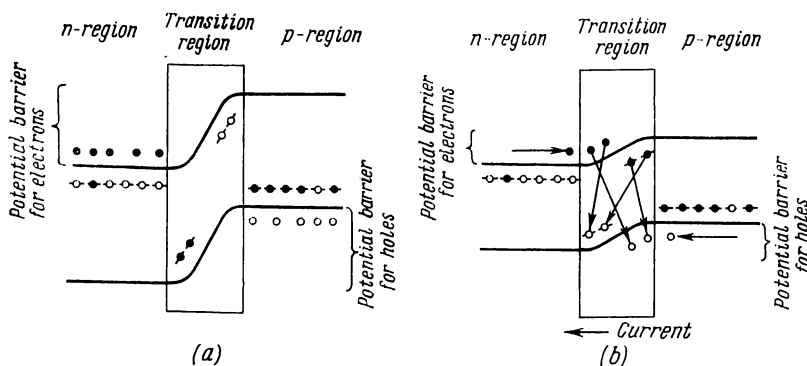


Fig. 322

supplied, and such a device would function as a refrigerator drawing heat from the crystal and surrounding medium. The entire emission propagates in the plane of the boundary layer. Two ends of a diode which are perpendicular to the boundary are polished until a resonance hollow is created. At heavy currents, emission becomes stimulated with all ensuing consequences as regards the sharpness of direction of polarisation and coherence.

Today we are in possession of a large number of semiconductor lasers. All of them belong to binary semiconductors combining the elements of columns II-VI, and also III-V of the Mendeleyev table. In accordance with the gap widths varying within the range of several electron volts, there are created semiconductor lasers covering the range of wavelengths from ultraviolet to remote infrared light.

#### Sec. 291. CHARGE DISTRIBUTION IN A NONUNIFORMLY HEATED BODY

Let us consider a rod along which a drop in temperature occurs. Different portions of the rod will be subject to different conditions, and this will affect the behaviour of the free electric charges. Where the temperature is higher the charges will have greater energy; moreover, the number of free charges may increase if electrons can pass from the filled band to the conduction band. Both effects tend to produce diffusion of free charges, which continues until a field which counterbalances the tendency to uniform distribution is created. A drop in potential will occur along the rod; negative charge is formed at one end of the rod and positive charge at the other. Each body has its own characteristic curve of potential drop as a function of temperature. The rate of fall of potential may be described by the

derivative of the potential with respect to temperature:

$$a = -\frac{d\phi}{dT}.$$

If a constant temperature difference is maintained between the ends of a rod, heat is transferred continuously through the rod. Heat is transferred by free charges, but current cannot flow in an open circuit. The continuous transfer of energy without the transfer of charge is achieved thanks to the different velocity of the charges moving from the hot end to the cold end, while the number of charges passing through a given cross-section per unit time is the same in each direction. If electrons are the carriers of current, an excess concentration of them is produced at the cold end of the rod. If positive particles or holes are the carriers of current, positive charge accumulates at the cold end. Thus, the sign of the potential difference will differ, depending on the sign of the current carriers.

Will the above effect occur in the case of a semiconductor, particularly one in which holes as well as electrons are the carriers of current? As a matter of fact, such bilateral diffusion may reduce to zero the potential difference of a nonuniformly heated body. However, the potential differences formed by positive and negative current carriers may not balance each other. This may occur as a result of a difference in mobility between electrons and holes, and also as a result of differences in their concentrations.

Certain difficulties occur in detecting potential differences in a nonuniformly heated conductor. The order of magnitude of such a potential drop is  $10^{-4}$  V/deg. This effect cannot be detected, of course, by forming a closed circuit of the conductor in the hope of measuring electric current. Such a closed circuit may be conceived of as divided into two halves: in one half a potential drop occurs and in the other a potential rise. In a uniform conductor, the magnitudes of these two potentials will be exactly equal; hence the emf we wish to measure will not be detected.

#### Sec. 292. THERMOELECTROMOTIVE FORCE

Electric current will flow in a wire ring consisting of two (or more) different materials if the junctions have different temperatures. This is the well-known thermoelectric effect, which has found broad practical application.

There are two possible reasons for the flow of a thermoelectric current. First, it is evident that the potential drops along the two wires due to temperature drop may differ if the values of the constant  $a = \frac{d\phi}{dT}$  differ for the two materials (we shall designate them as I and II).

Thus, the potential differences  $\int_{T_1}^{T_2} a_I dT$  and  $\int_{T_2}^{T_1} a_{II} dT$  generally are not equal.

This alone would be sufficient for an emf equal to the difference between these voltages to arise in the wire ring.

The second reason for thermoelectric current lies in the fact that contact potential quite probably depends on temperature. If the two junctions are placed at different temperatures, their contact potentials may differ. Again, this condition alone would be sufficient for a net potential difference to exist in the closed circuit and, hence, for a current to flow.

Taking both phenomena into account, we may express the thermoelectromotive force as the sum of the voltage drop in the first wire, the jump in potential from the



first wire to the second, the potential drop in the second wire, and the jump in potential from the second wire to the starting point of the circuit:

$$\mathcal{E} = \int_{T_1}^{T_2} a_I dT + [\varphi_{II}(T_2) - \varphi_I(T_2)] + \int_{T_2}^{T_1} a_{II} dT + [\varphi_I(T_1) - \varphi_{II}(T_1)].$$

To simplify the above expression, let us write the difference  $\varphi_{II}(T_2) - \varphi_{II}(T_1)$  in the form

$$\int_{T_1}^{T_2} \frac{d\varphi_{II}}{dT} dT$$

and the analogous second difference in a similar form. Now, the formula for the emf assumes the form

$$\mathcal{E} = \int_{T_1}^{T_2} \left( a_I - \frac{d\varphi_I}{dT} \right) dT - \int_{T_1}^{T_2} \left( a_{II} - \frac{d\varphi_{II}}{dT} \right) dT.$$

Thus, we have succeeded in expressing  $\mathcal{E}$  in the form of a difference between two quantities, each of which is characteristic of a given body. Quite often the term "thermoelectromotive force" is used to refer to the emf per degree:

$$\alpha = a - \frac{d\varphi}{dT},$$

rather than to the above integral. This quantity is a fundamental characteristic of the thermoelectric properties of a body. It is not an invariable constant, for it may depend on thermodynamic conditions, including the temperature. However, for many bodies this dependence is not well defined.

By measuring the thermoelectromotive force, we may determine the difference between the above quantities, but we cannot determine  $\alpha$ . However, by forming different pairs of conductors and semiconductors, we are able to determine the value of  $\alpha$  relative to a material taken as a "base". Thus, materials may be arranged in a series in accordance with their thermoelectromotive forces. For reasons which are quite understandable in view of what has been said, a thermoelectromotive force series does not coincide with the corresponding contact potential series.

Let us list the thermoelectromotive forces of several metals with respect to platinum. If a given metal is joined to platinum and one junction is held at 0°C while the other is held at 100°C, an emf arises in the closed circuit:

Antimony	+ 4.0 mV
Iron	+ 1.9 "
Copper	+ 0.75 "
Nickel	- 1.5 "
Constantan	- 3.4 "

The positive sign indicates that at the 0°-junction current flows from the given metal to the platinum.

Using a table of values of the constant  $\alpha$ , one can calculate the thermoelectromotive force occurring for a given temperature difference from the expression  $\mathcal{E} = (\alpha_1 - \alpha_2)(T_1 - T_2)$ . This is the expression for a thermal element consisting of two metals or two semiconductors the current carriers of which have the same sign. In this case, the potential differences arising in the two branches of the circuit are in phase opposition, and the resulting emf is equal to the difference

between the effects of the two conductors forming the circuit. However, the picture changes when the circuit is formed of two semiconductors, one of which possesses hole conductivity and the other electron conductivity. In a *p*-type conductor, holes move toward the cold junction and electrons toward the hot junction. In an *n*-type conductor, electrons move toward the cold junction. The two effects reinforce each other and the formula assumes the form.

$$\mathcal{E} = (\alpha_1 + \alpha_2) (T_1 - T_2).$$

This fact is of great practical significance.

#### Sec. 293. LIBERATION OF HEAT IN ELECTRICAL CIRCUITS

Joule heat is liberated in a conductor in which current flows. The displacement of charges along a body is accompanied by two other thermal effects.

The first of these, the Peltier effect, consists in the following. If electric current passes through a junction between two bodies, heat proportional to the current strength is released or absorbed at the junction:

$$Q = \Pi I,$$

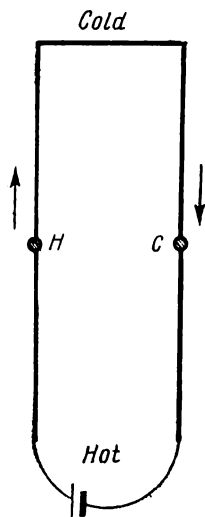


Fig. 323

where  $\Pi$  is a proportionality constant. A remarkable feature of this effect is that the sign and magnitude of the heat changes when the direction of the electric current changes, i.e., depending on this direction, a particular junction will release or absorb heat. This was demonstrated by Lenz more than a century ago. A drop of water was placed in a recess at the junction between an antimony rod and a bismuth rod. Then, by passing current in one direction, he showed that the drop freezes. When the current was reversed, the drop melted.

The second effect occurs in any uniform conductor which is heated nonuniformly. Assume, for example, current flows along a rod, one end of which is maintained at one temperature and the other end at another. In such a conductor additional heat, proportional to the first power of the current strength (not to the second power as in the case of the Joule effect), will be liberated. To detect this effect, we must, of course, reduce the Joule heat to a minimum.

This effect, which was predicted by the British physicist Thomson on the basis of thermodynamic considerations, may be demonstrated in the following manner. Included in the current circuit are two bars which are placed parallel to each other as shown in Fig. 323. The ends of these bars, maintained in pairs, are held at different temperatures. It would seem, in view of the symmetry of the arrangement, that symmetrical points of the bars should have the same temperature. However, in one bar the current flows from the hot end to the cold end and in the other from the cold end to the hot one. Owing to the Thomson effect, corresponding points of the two bars do not have the same temperature. A point of the bar in which the current flows from the hot to the cold end will be hotter than the corresponding point of the other bar.

The quantity of heat liberated per second in a segment of length  $dx$  may be written in the form

$$dQ = \tau I \frac{\partial T}{\partial x} dx,$$

where  $\tau$  is a proportionality constant. The greater the temperature gradient, the greater the quantity of heat. Three effects exist simultaneously in a thermoelectrical circuit: the appearance of a thermoelectromotive force, the Peltier effect and the Thomson effect. On the basis of the principles of thermodynamics, it can be proved that these three processes are interconnected. This requires no proof for weak currents. Since the thermoelectric effects are proportional to the first power of the current and the Joule heat to the second, the Joule heat is negligible in such cases.

*Example.* The ends of a rod of sodium ( $\tau = -8.5 \times 10^{-6}$  V/K and  $\rho = 5 \times 10^{-6}$  ohm cm) of 10-cm length and 5-mm<sup>2</sup> cross-section are maintained at temperatures of 300 K and 310 K. When a current  $I = 0.5$  mA flows from the hot to the cold end of the rod, the heat liberated in the conductor per unit time due to the Thomson effect is

$$Q_T = \tau I \frac{\partial T}{\partial x} l = -8.5 \times 10^{-6} (-5 \times 10^{-4}) \times 1 \times 10 = 4.24 \times 10^8 \text{ J/sec.}$$

The minus sign preceding the current indicates that the current flows in the direction of decreasing temperature. The heat liberated in the conductor per unit time due to the Joule effect is

$$Q_J = I^2 R = (5 \times 10^{-4})^2 \times 5 \times 10^{-6} \frac{10}{5 \times 10^{-2}} = 2.5 \times 10^{-10} \text{ J/sec,}$$

i.e., about  $\frac{1}{200}$  of the value of the Thomson heat.

Thermodynamic analysis shows that the coefficients  $\alpha$ ,  $\Pi$  and  $\tau$  are interrelated as follows:  $\tau = \frac{\partial \Pi}{\partial T} - \alpha$  and  $\alpha = \frac{\Pi}{T}$ . Substituting  $\alpha T$  for  $\Pi$  in the first relation, we obtain:  $\tau = T \frac{\partial \alpha}{\partial T}$ . The absolute value of  $\alpha$  may be determined from these equations.

The Peltier and Thomson effects have the same physical basis as thermoelectromotive force. In the final analysis, a thermoelectromotive force arises from the fact that heat flow transfers electric charges. Here, however, we are dealing with phenomena in which a flow of electric charges transfers heat.

#### Sec. 294. APPLICATION OF THE THERMOELECTRIC EFFECT

The opportunities for the application of thermocouples as generators of electrical energy have increased considerably in recent times. Metal thermocouples have a coefficient of efficiency of the order of 0.5 per cent, but that of a semiconductor thermocouple consisting of a hole segment and an electron segment has already reached as much as 7-8 per cent. The low efficiency results from irreversible losses in the form of the Joule heat. If  $R_0$  is the resistance of the internal portion of the circuit and  $R$  that of the external circuit, the power delivered to the external resistance (useful power) will equal  $\frac{\mathcal{E}^2 R}{(R + R_0)^2}$  for any electrical circuit; here,  $\mathcal{E}$  is the emf. Substituting the value of the thermoelectromotive force, we obtain for the power of a thermocouple the expression

$$\alpha^2 (T_1 - T_2)^2 \frac{R}{(R + R_0)^2}.$$

The electromotive force of a thermocouple is of the order of several tenths of a volt. To obtain a voltage of 120 V, for example, thermocouples are connected in series like a battery. If heavy currents are required, thermocouples must be connected in parallel.

Another important application of the thermoeffect, which has also become possible as the result of the development of semiconductor engineering, is the employment of semiconductors as a refrigerator.

The application of the thermoelectric effect to the measurement of temperature is well known and need not be discussed.

An important and long known field of application of the thermoeffect is in the detection of very small amounts of heat. The opportunities in this field have increased still further as a result of the fact that semiconductors yield large thermoelectromotive forces. For such purposes, thermocouples connected in series—so-called thermopiles—are used. Every other junction of a thermopile is cooled and the alternate ones are heated. Thermopiles are used to measure power levels as low as several ergs per second. However, it is possible to lower this limit still further, i.e., to several tenths of an erg per second. This is achieved by means of vacuum thermocouples, the thermal losses of which are reduced to a minimum.

#### Sec. 295. MICROELECTRONIC CIRCUITS

Valves and transistors are applied not only for transmitting radio waves. Therefore the demarcation line drawn between two branches of applied physics: radio engineering and electronics is of a conventional character. The task of setting forth new material and introducing various classifications has turned out to be still more difficult with the birth of microelectronics.

The appearance of transistors which made it possible to solder electronic and other circuits was, undoubtedly, a revolutionary event. It caused an abrupt reduction of dimensions and weight of various devices, made them cheaper and more efficient.

But the next discovery which resulted in passing from electric circuits whose elements had to be interconnected through wires to electric circuits “drawn” on a small piece of silicon (only several millimetres in size) deserves a special name. This passage began some ten-fifteen years ago and is far from its completion. Many researchers assert that it has led to an intellectual revolution, since it allowed man to entrust machines with almost all functions which long since have been regarded as the privilege of human brain.

Microelectronics is the subject which teaches us how, with the aid of miniature circuits placed on a silicon chip, we can carry out mathematical and logical operations, transform one form of information into another, and transmit it over a distance, as well as store it for an arbitrarily long time.

Microelectronic devices have neither capacitors, nor resistors as separate objects. A semiconductor substrate itself (silicon or germanium) makes it possible to create electrical resistance between *p*-type and *n*-type areas. As far as capacitance is concerned, it is not difficult to obtain it. To this end, one should master the technique of depositing an insulating layer on the semiconductor surface to be covered then with a conducting layer. Unfortunately, for the time being we have no method for “constructing” inductance inside the silicon. However it may be, we can do without self-inductors and this shortcoming is compensated with usury by the possibility to make various diodes and triodes in the semiconductor body. The possibility of creating transistors of various types in silicon is, undoubtedly, the most important peculiarity of microelectronics.

First we shall describe the principle of operation of microelectronic transistors, and then add a few words about their manufacturing.

Figure 324 shows a diagram of the so-called bipolar transistors fabricated in a single crystal of silicon through a series of operations that require access to only

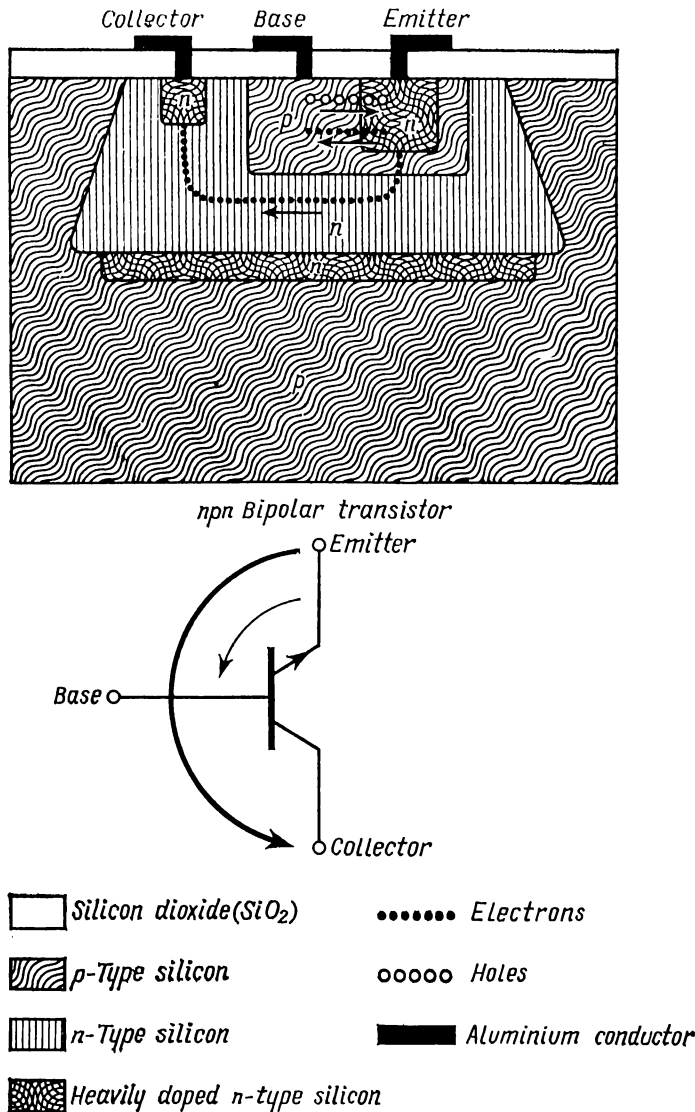


Fig. 324

one surface of the silicon chip. In the example represented here the entire chip is doped with *p*-type impurity, then islands of *n*-type silicon are formed. Smaller *p*-type and *n*-type areas are created within these islands in order to define the three fundamental elements of the transistor: the collector, the base, and the emitter. In an *npn* transistor a positive voltage is applied to the base and the collector, and as a result holes flow from the base to the emitter and electrons are injected by the emitter into the base. Many of the injected electrons, however, migrate all the way through the base to reach the collector, and this emitter-to-collector current can be much larger than the emitter-to-base current. The device exhibits gain because a small signal applied to the base can control a large signal at the collector.

Current flows through aluminium conductors deposited over an insulating layer of silicon dioxide. Some areas of  $n$ -type silicon are heavily doped to improve their conductivity. The large  $n$ -type islands are required to isolate the transistors. Because charge carriers of both polarities participate in the operation of these devices they are called bipolar transistors.

We are not going to describe the design of a  $pnp$  transistor. It differs somewhat from the  $npn$  type, but the principle of operation is the same.

Widely used today are transistors in which only kind of charge carrier is active in a single device. They are called field-effect transistors. Those in which current

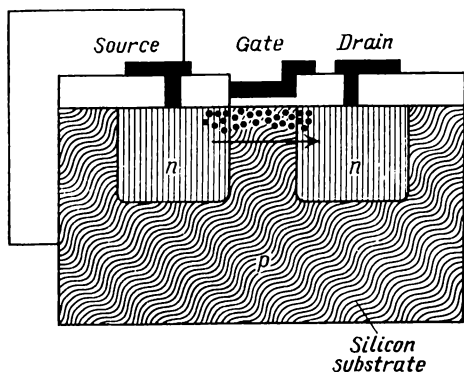


Fig. 325

is carried by electrons are termed  $n$ -transistors, since they employ two  $n$ -islands (called the source and the drain) in  $p$ -type silicon substrate. Over the channel, along which electrons flow, between the source and the drain is a metal electrode, the gate, that is prevented from making contact with the semiconductor by a thin layer of silicon dioxide (see Fig. 325). But this electrode, together with silicon, form a kind of capacitor. Therefore, an electric field applied to the middle electrode can essentially affect the flow of electrons. The source and the silicon substrate are usually linked with an outside conductor and are kept at zero voltage. A positive

charge is supplied to the drain. There is no current between the source and the substrate, a negligible reverse current flows between the drain and the substrate. Applying the plus to the middle electrodes, we shall make all electrons to gather in a thin layer close to the surface of the crystal. Figuratively speaking, we may say that the  $p$ -type area is inverted in this layer. As a result, a channel of  $n$ -type is created between the source and the drain along which heavy currents can flow. The same as in bipolar transistors, we get here the possibility to amplify small signals.

Thousands of such transistors can be deposited on a single piece of silicon a square millimeter size. And, according to technologists' opinion, this is not the limit. Miniaturising of computers and various radioelectronic devices affords a phenomenal opportunity for science and technology.

#### Sec. 296. TECHNOLOGY OF MANUFACTURING MICROELECTRONIC CIRCUITS

The principles underlying the production of tiny crystals of silicon with electric circuits deposited on them were worked out some ten years ago, and since that time have remained unchanged. But the manufacturers achieve still new successes each year owing to arrangement of a greater number of elements on the same area of a crystal surface. Roughly speaking, this number is doubled annually during the last five years.

An electric circuit consisting of about ten thousand elements is contained on a small area with the linear dimensions about two millimetres. Each element has a three-dimensional structure, which means that one has to create several layers on a tiny area of silicon surface; some of these layers lie inside a silicon chip, others being deposited on its surface.

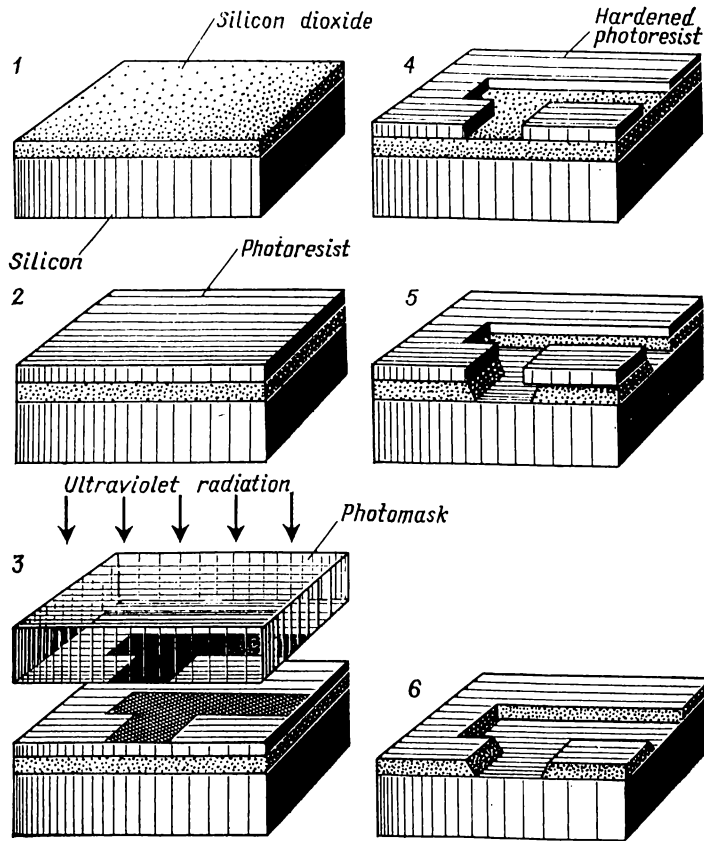


Fig. 326

You will grasp this idea looking at Fig. 326 which illustrates the process of photolithography.

Photolithography is the process by which a microscopic pattern is transferred from a photomask to a material layer in an actual circuit. In this illustration a pattern is shown being etched into a silicon dioxide layer (colour) on the surface of a silicon wafer. The oxidised wafer (1) is first coated with a layer of a light-sensitive material called photoresist (2) and then exposed to ultraviolet light through the photomask (3). The exposure renders the photoresist insoluble in a developer solution; hence a pattern of the photoresist is left wherever the mask is opaque (4). The wafer is next immersed in a solution of hydrofluoric acid, which selectively attacks the silicon dioxide, leaving the photoresist pattern and the silicon substrate unaffected (5). In the final step the photoresist pattern is removed by means of another chemical treatment (6).

Using various types of the mask and a series of chemicals we can dope necessary areas of the crystal surface with *p*-type and *n*-type impurities. To this end, treatment with ion beams is carried out. Polycrystalline silicon is utilised to create conducting areas.

We are not going to enter into details of this technology, but we hope that the reader is now familiar with the general idea.

The inside of a wafer-fabrication facility must be extremely clean and orderly. Because of the smallness of the structures being manufactured even the tiniest dust particles cannot be tolerated. A single dust particle can cause a defect that will result in the malfunction of a circuit. The air is continuously filtered and recirculated to keep the dust level at a minimum. The quality of material (the absence of crystal defects) is another important factor. A scratch even one micrometre long can destroy an electric circuit. It is impossible to remove a scratch or to find a dust particle, that is why a faulty element should be simply rejected. Flaw detection is carried out with the aid of scanning electron microscopes.

It is quite clear that with the aid of photolithography it is impossible to deposit (on a small crystal) elements whose size is less than the length of a light wave. But technologists intend to make use of X-ray lithography. This yields unlimited possibilities for further miniaturisation, since the wavelength of X-rays are of the same order as the distance between atoms. But the reader should not think that there are no difficulties here. The masks are to be made from metal, it is difficult to adjust them in a precise manner. Also, exposure must be accomplished in vacuum, since X-rays are absorbed by air.

Thus, we are now still at the stage of hopeful experiments.

It is possible to use lithography with the aid of electron streams. This method is still very expensive, but promises rich possibilities.

#### Sec. 297. MICROPROCESSORS

Microprocessor is the name given to the 'heart' on an electronic device. Its function is to accomplish arithmetical and logical operations. It usually represents a single small piece of silicon containing hundreds or even thousands of transistors, resistors, and capacitors connected in a proper way. Today, a microprocessor is some five millimetres in size!

A microprocessor connected with several dozens of other chips which serve as a memory, input, output, and other devices form a microcomputer measuring a page of a book.

It is also possible to manufacture a single-chip microcomputer. Figure 327 shows a photograph of such a device. You are given this picture, of course, not to gain an understanding of the represented circuit, but just to strike your imagination! This single-chip microcomputer (which is an electronic computing machine!) measures only 5.6 by 6.6 millimetres, but it is a complete general-purpose digital processing and control system in one large-scale integrated circuit. The device combines a microprocessor, which would ordinarily occupy an entire chip, with a variety of supplementary functions such as program memory, data memory, multiple input-output (I/O) interfaces and timing circuits. The program is stored in an erasable and reprogrammable read-only memory, which has a capacity of 8,192 bits (binary digits)\*. The program is erased by exposing the circuit to ultraviolet radiation, which causes the electric charges stored in the memory device to leak away, after which a new program can be entered electrically.

The functions of the microprocessor are: to receive and to record information in the form of a sequence of unities and zeros, to store it until it is needed, to carry out logic and arithmetic operations on this information according to the instruc-

---

\* The bit is a unit of information. The number of bits is equal to the number of 'sequence of zeros and unities with the aid of which a number is written in the binary system. By the way a dictionary containing 10,000 words is more than sufficient to have a good command of a foreign language! Now the reader will understand that the above number is not at all small.



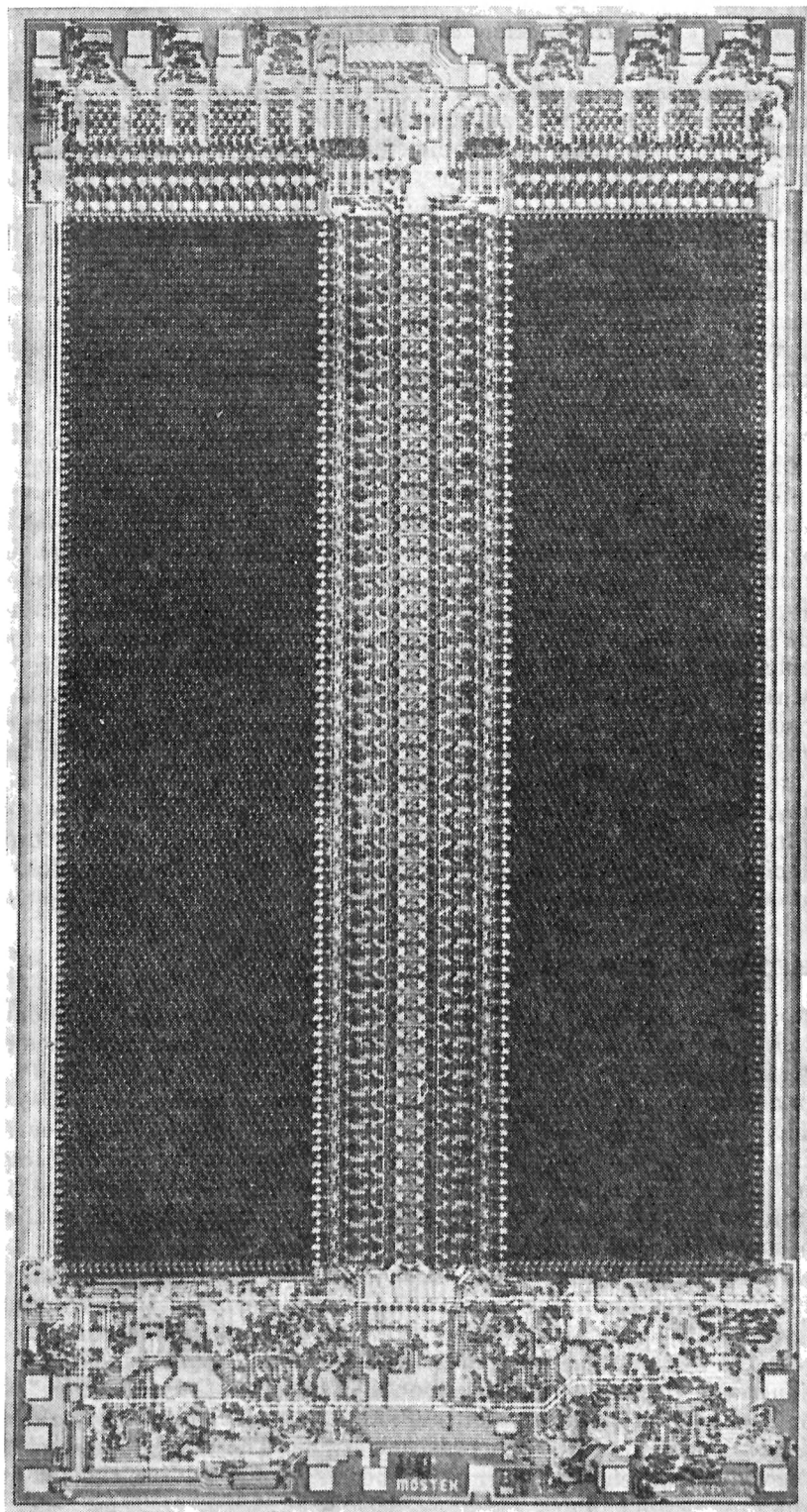


Fig. 327

tions received (programs), and to serve out the result to the user (on an electric typewriter, or on the screen of an electron tube, or in some other way).

Microprocessors differ from microcomputers in many respects: by manufacturing technology, design, capacity, and type of use. Their classification is rather complicated even for specialists.

Of course, the reader is already familiar with cheap computing machines each of which can be "installed" on our palm. Such a computer is capable of satisfying the needs of any student and engineer. Trigonometric functions, any powers, any roots, conversion from radians to degrees and hundreds of other applications can be served by a palm-size computer fitted with a small memory. Some twenty years ago a room would be required to accommodate such a multi-operation computing machine!

Applications of microprocessors are practically unlimited. Soon they will occupy a due place in automobiles to check the operation of the engine and to assist the driver in driving his automobile. In an office, side by side with a typewriter, we shall see a microcomputer which will store and put out with lightning speed any necessary information concerning the personnel of a large enterprise. One microcomputer will successfully replace the personnel department. In a couple of years control of any production and quality inspection of manufactured goods will be completely entrusted to microcomputers. At home a microcomputer will become as ordinary a thing as a TV-set, or a refrigerator. It will be able to check and receive any data at your will. Instead of eighty volumes of Great Soviet Encyclopaedia, it will be sufficient to be in possession of a small microcomputer. In modern airplanes microprocessors direct the flight, condition the air in a pilot's cabin, fulfil the requests of passengers, follow the operation of engines, control the positions of wing- and tail-flaps.

In the future many brilliant successes which are difficult to be comprehended even by science-fiction writers will be achieved by microelectronics, this new branch of applied physics.

#### Sec. 298. ELECTRONIC COMPUTERS

Hitherto it was unusual to dedicate a single section to electronic computers in books on physics. This is unfair, since computing devices and machines constitute a separate section of applied physics. A chapter about computers has the right to exist in a book on physics and in a course of lectures in physics no less than have the sections dedicated to the physical principles of electrical engineering, radio engineering, thermal engineering, etc. Naturally, since we speak of a huge branch of industry, the aim of a book on physics is, as usually, to set forth the basic principles and ideas. It is absolutely necessary for everybody educated to the slightest extent to study them thoroughly.

The computer is rightfully called an artificial brain, since it can accomplish a number of operations and even more rapidly and precisely than a human brain. And, in general, there is no problem which would puzzle the computer.

The principal devices of a computer are: an "input", that is a device receiving information (with respect to man we would say "task", "visual image", "auditory impression", and so on), and an "output", that is a device which yields the solution of a problem in the form of words or actions. The machine has a "memory", that is a device in which the information necessary for solving the set problem is stored.

The most important feature of the computer is that one and the same machine can solve a numerous (in principle, any) class of problems which are dictated by

the "program", i.e. by the formulation of a problem using the language and technical means predesigned for a given computer.

The quickness with which the computer copes with a problem is, perhaps, the leading reason why we assert that computers have ushered in an intellectual revolution. It can easily cope with problems which are considered to be unsolvable for the human brain since their solution would take inconceivably much time.

This circumstance leads to a qualitative jump since we are enabled to deal simultaneously with vast information the knowledge of which helps us to draw a final conclusion.

The idea that a machine can be charged with calculations is not a new one. As back as in the seventeenth century the great French geometer, probabilist, physicist, and philosopher Blaise Pascal (1623-1662) invented and constructed first calculating machine which employed the same principle of operation as the modern computers do, that is the binary number system. In the middle of the last century Charles Babbage designed a machine capable of carrying out complicated calculations. He was assisted by lady Lovelace, daughter of the great English poet George Gordon Byron. She was evidently the first to suggest the idea of changeable "programs" of computation. In 1840 she published her program for calculating the so-called Bernoulli numbers.

But the first practical instructions as how to create a computer (which resembled a modern computing machine not to a greater extent than Stephenson's locomotive resembles a jet plane!) were given only thirty years ago.

It is not difficult to name a number of simple physical devices which may be found in two states. For instance, an electron tube can be "locked" (shut) or open for an electron flow. The same with a transistor. An iron body can be readily brought to a demagnetised or magnetised state.

Such a device is said to store "one bit of information". The open and closed states may be denoted as plus and minus, or 0 and 1, or A and B. Let us take zero and unity, recalling that in the binary system their sequence represents any number. Suppose we have a machine consisting only of five such elements. With their aid we can express  $2^5$ , i.e., 32, numbers: 00000, 00001, 00010, . . . , 11111. This is quite sufficient to represent the entire alphabet.

With the computer elements fixed in a definite state, we are capable of writing down any information. We may control these "notations" by switching the computer from state 0 to state 1 and then back to 0.

This just enables us to perform logical and mathematical operations of any degree of complexity if the computing machine contains a sufficiently large number of elements, i.e., triodes and diodes.

## Sec. 299. ELECTRONIC ARITHMETIC

Any binary number in the binary system is a sequence of unities and zeros. This fact leads to such a situation that practically all arithmetic operations are reduced to addition. The rules for binary addition are just the same as for decimal numbers we are used to. But, please, do not forget about the figure "which should be carried".

Consider the binary addition of the numbers 9 and 5, i.e. 1001 and 0101. To begin with, 1 and 1 yield 2. Since the number system is binary, 0 is written and 1 is carried. Now we have to add 0 and 0; the result would be 0 if we had not got 1 carried in our mind. The addition of the figures occupying the third column yields 1, and the same result is obtained from adding the figures 1 and 0 forming

the fourth column. Thus,  $+1001$ , i.e., the number 14 in the binary system (1 occupying the first place means 8, that is  $2^3$ , 1 in the second place represents 4, that is  $2^2$ , 1 in the third place  $2^1$ , that is 2, and, finally, 0 occupying the last place means absence, that is  $2^0$ ).

We are not going to show that multiplication is reduced to addition, subtraction to addition of a "supplementary" number, and division to repeated subtraction. This is arithmetic for entertainment, but yet not physics.

The binary number system is quite suitable for a computer to formulate logic conclusions. It is assumed that 1 means correctness of a judgement and 0 its falseness.

Mathematical logic (created long before computers by the pioneering British logician George Boole) shows that three logic functions, called "not", "and", "or", are sufficient for a logic analysis. "Not" is the simplest of them and is symbolised

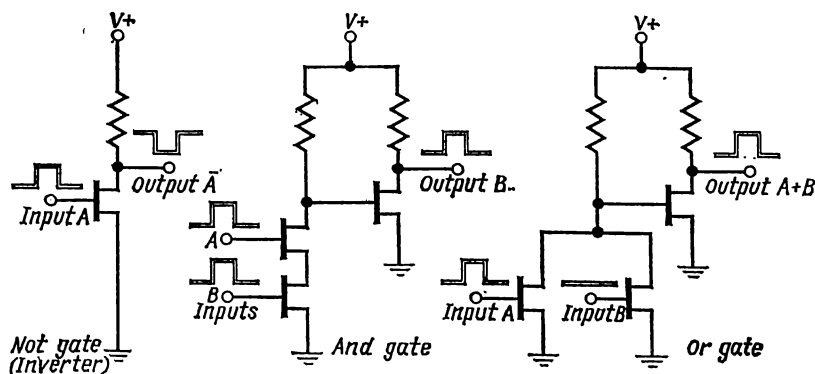


Fig. 328

by a change of 1 for 0, or vice versa; "and" means the appearance of 1 at the "output" only if all the inputs into one "cell" are marked with 1, and, finally, "or" is occurrence of 1 at the output if there is 1 at least at one of the many inputs. For instance, the operation "and" means the choice by a machine out of the set of objects possessing three properties (say, shape, colour and weight), only those which are black, round, and heavier than one kilogram at the same time. And "or" means the choice of objects which are black, or round, or heavier than one kilogram.

Arithmetic and logic operations are carried out by elements or cells or, as they are called more precisely, by electronic logic gates of a computer.

Figure 328 presents the diagrams of these three types of gates which enable the computer to cope with, we should say, fantastic assignments.

Electronic logic gates evaluate arithmetic and logic expressions in which binary values are represented by voltages. By convention, a binary 1 is represented by a high voltage and 0 by a low voltage. The gates shown here are constructed from metal-oxide-semiconductor transistors. The simplest is the "not" gate, or inverter. When the input of this gate is in the low state, the transistor does not conduct and only a negligible current flows from the supply voltage ( $V+$ ) through the resistor and transistor to ground. As a result there is little voltage drop across the resistor and the output is in effect connected to the supply voltage. When a high signal is applied to the input, the transistor conducts and the comparatively large current

flowing through the circuit produces a considerable voltage drop across the resistor. The output voltage is now near ground and is therefore in the low state. The "and" gate has two input transistors connected in series current flows through them only when both receive a high signal simultaneously. In order to restore the proper polarity of the signal, the output of the two series transistors is followed by an inverter. In an "or" gate the input transistors are arranged in parallel, so that a high signal applied to either one of them results in conduction. Again, an inverter is required to change the polarity of the output.

It is understood here that computers use "field" transistors.

Let us now consider the process of addition of two numbers written as sequences of 0's and 1's. All possible additions are exhausted by the following simple table which is known as a truth table.

A	B	C	Sum bit	Carry bit
0	0	0	0	0
0	0	1	1	0
0	1	0	1	0
0	1	1	0	1
1	0	0	1	0
1	0	1	0	1
1	1	0	0	1
1	1	1	1	1

Truth tables for binary addition give the calculated values of the sum bit and the carry bit for all possible values of the input. The inputs are the two bits to be added (A and B) and the carry bit (C) from the next lower power of 2. The rules for addition specify that the sum bit is a 1 if exactly one of the inputs is a 1 or if all three are 1's; otherwise the sum bit is a 0. The carry bit (for the next higher power of 2) is a 1 if at least two of the inputs are 1's.

By simply ganging together electronic logic gates, a one-bit binary adder could be built from about 50 transistors and almost as many resistors. Even in its most simplified form the adder circuit (see Fig. 329) is a logic array of considerable complexity.

Such a binary adder provides an array of logic gates for each combination of input bits (A, B, and C) that requires an output. The four rows of gates at the bottom calculate the sum bit; the three rows at the top calculate the carry bit for the next higher power of 2. As an example, suppose both of the bits to be added (A and B) are 0's but the carry bit (C) from the preceding column is a 1; the above given truth table indicates that this combination of inputs must yield a sum bit of 1 and a carry bit of 0. The sum bit is generated in the fourth row from the bottom, where A and B are applied to an "and" gate in inverted gate, so that both appear as binary 1's. The output of this gate, which is a 1, is combined with the carry bit in a second "and" gate, and the 1 output is passed through a series of "or" gates to the output of the adder. None of the gates for the calculation of the carry bit responds to this combination of inputs, and so the carry bit is a 0; if two or more of the inputs were 1's, then a carry bit of 1 would be issued. A binary adder requires a logic array like this one for each bit of the numbers to be added, with the carry-output line of each stage connected to the carry-input line of the next stage.

It is easy to calculate that the adder consists of 51 transistors. Actually an adder need not be that complex. Refinements to the design that allow one transistor to serve more than one function reduce the number of elements to 17 transistors and four resistors, or merely 21 transistors.

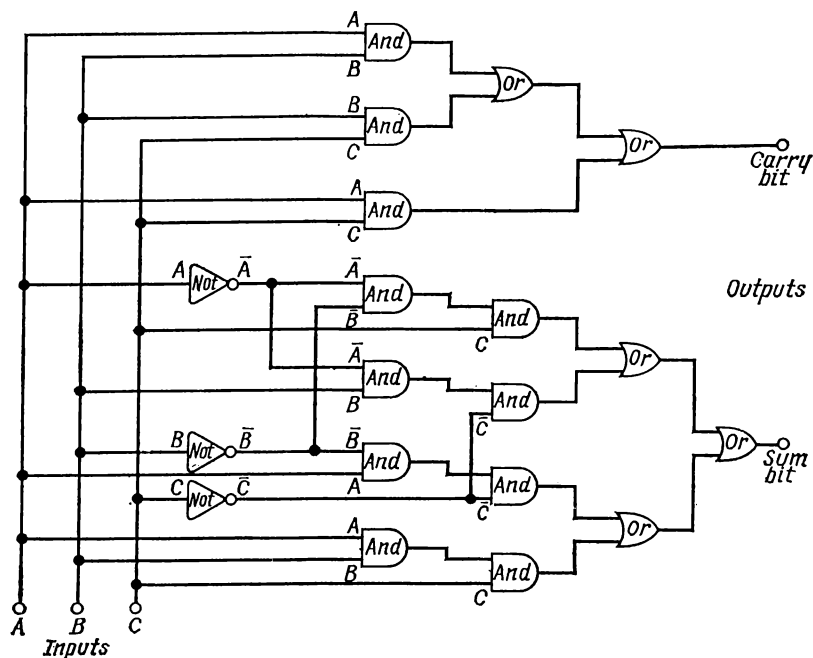


Fig. 329

Combining adders of the described type (i.e., arranging them so that the output of one adder serves as the input of another), we can carry out arbitrarily complicated arithmetic and logical operations.

The logical and arithmetic manipulation of binary numbers is not the only function of the basic gate circuits. Another essential function in computers is the storage of information.

#### Sec. 300. ELECTRONIC MEMORY

A modern small pocket-size computer is capable of storing about 100 bits of information, i.e., it has 100 cells, a fraction of which being in state 0, the rest of them in state 1. Thus, in this case 1's and 0's can be arranged in  $2^{100}$  ways. Then what can be said about a large computer capable of manipulating information reaching one trillion bits!

Like in man, a computer can memorise for a long and for a short period of time. The long-time memory stores, so to say, absolute information (say, the data on the lengths of all the rivers on the Earth or the coordinates of stars in the sky), whereas the short-time memory device "memorises" numbers or other data necessary for a short period of time during which a certain logical or mathematical operation is being carried out.

Information is stored by a gramophone record, magnetic tape, etc. But we are speaking of a computer memory in which various kind of data are stored in cells

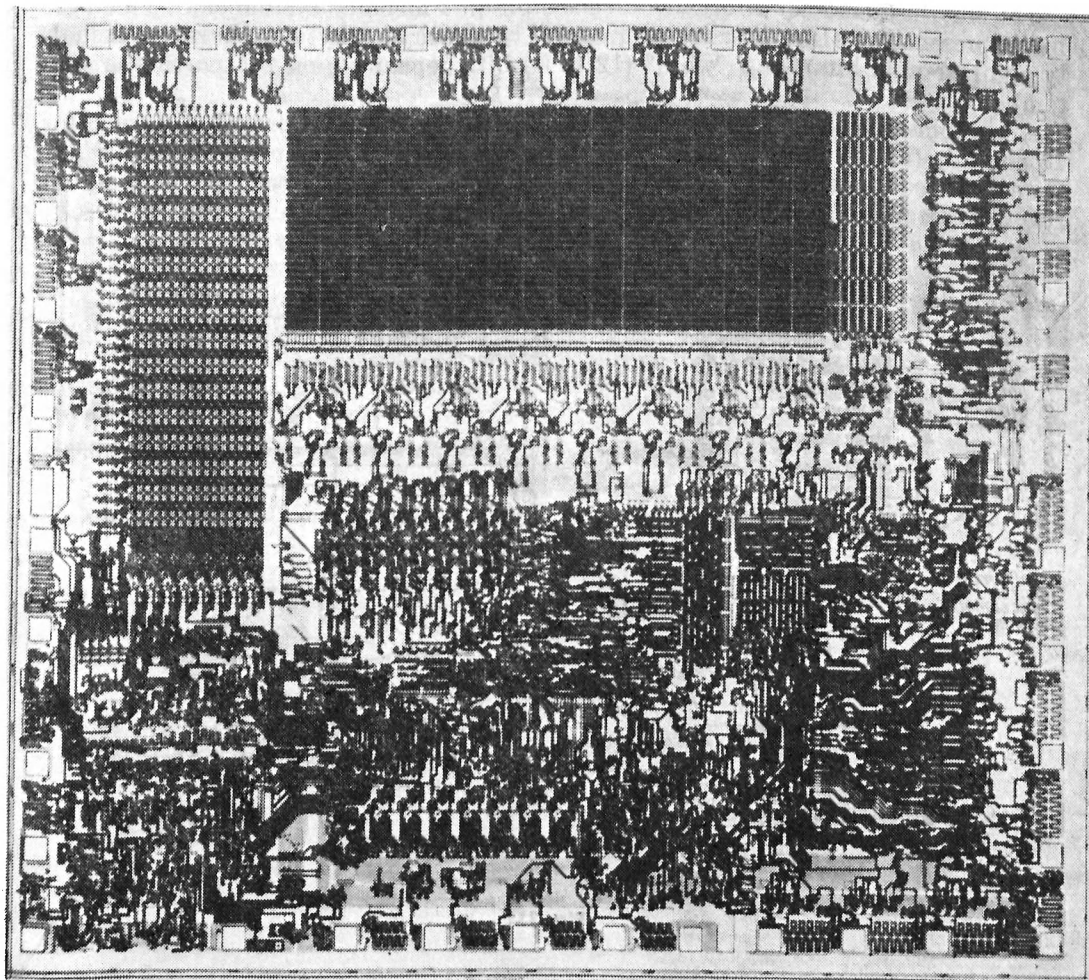


Fig. 330

in the form of 0's and 1's. In most cases we are interested in a readily-accessible memory whose information can be, if necessary, "erased", and new data recorded. It goes without saying that for computer operation of great importance is not only the total number of cells, but also the time during which one bit of information can be "read out" from any place of the record.

A simple one-transistor storage cell is widely used today in the most economical random-access read/write memory devices. Information is stored as an electric charge on a small capacitor. The value of the capacitance is of the order of 50 femtofarad ( $50 \times 10^{-15}$  farad). For the representation of binary information two different levels of stored charge are needed. A binary 0 might be represented by zero charge and a binary 1 by a charge of 500 femtocoulombs (equivalent to 10 volts on the storage capacitor). Although this may seem like a tiny amount of charge, it is the charge on three million electrons. Reliable binary storage should ultimately be attainable with a charge 1,000 times smaller.



The smallest block of information accessible in a memory system can be a single bit (represented by 0 or 1), a larger group of bits such as a byte or character (usually eight or nine bits), or a "word" (12 to 64 bits depending on the particular system).

A random-access memory (RAM) circuit shown in Fig. 330 provides storage for 16,384 bits (binary digits). Each bit is held in a single-transistor storage cell. The time required to write one bit in any arbitrary location or to read it out is about 200 nanoseconds. The RAM chip shown here measures 2.8 by 5.1 millimetres.

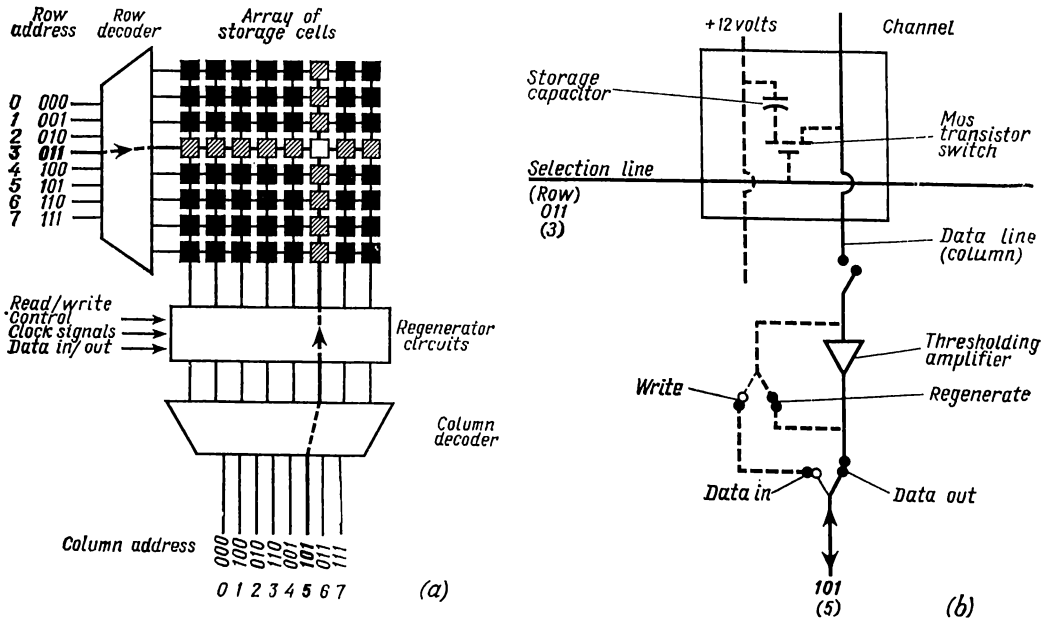


Fig. 331

Random-access memories are usually organised in rectangular arrays of rows and columns. The diagram of Fig. 331 (a, b) (left) shows an eight-by-eight array for storage of 64 bits, one bit being stored in the cell at each intersection. To specify a particular memory location three binary digits are needed to indicate the row and another three to indicate the column. In this example row address 011 (binary for 3) and column address 101 (binary for 5) specify the memory location 3.5. (The locations start with 0.0 at the upper left and end with 7.7 at the lower right, specifying 64 locations in all.) The organisation of a single one-transistor memory cell of the array is depicted at the right. Binary information is stored as a charge on a small capacitor. For example, zero charge might represent a binary 0 and a charge of  $500 \times 10^{-15}$  coulomb might represent a binary 1. When one of the selection lines, or rows, in an array is activated (here it is row 3), it turns on all the transistor switches connected to it. The transistor functions as an on-off switch to connect the storage capacitor to its particular data line, which corresponds to a column in the array. The simultaneous activation of a row and a column identifies the cell selected for reading or writing (here cell 3.5). Because the storage capacitor loses charge both by being read and by leakage it must be regenerated periodically, usually once every two milliseconds. The regenerated charge, which is



supplied by the thresholding amplifier, is returned to the capacitor by the closing of the switch in the data line. All the switching in this type of memory device is accomplished by transistors.

\*       \*

The current period is characterised by a rapid development of computer technology; in particular, we have seen a marked proliferation of computers (by several orders of magnitude) over the last decade. This is accompanied by the use of computers and numerical methods for a rapidly expanding range of problems.

Characteristic of the present period is the great expansion in applications of mathematics, which is largely due to the advent of new and powerful computing machines. With the introduction of the program-controlled computer, the speed of computation has increased over the past thirty years from 0.1 operation per second in the case of hand computation to 3,000,000 operations per second on modern serially produced computers—roughly a  $5,000^2$ -fold increase.

# APPENDIX

TABLE OF ELECTRODYNAMIC FORMULAS

	Physics formulas	Elec. eng. formulas
Electric field intensity	$E = \frac{F}{q}$ $\frac{\text{dyne}}{\sqrt{\text{dyne/cm}}}$	$E = \frac{F}{q}$ $\frac{\text{N}}{\text{V/m}}$
Surface charge density	$\sigma = \frac{dq}{dS}$ $\frac{\sqrt{\text{dyne/cm}}}{\text{cm}^2}$	$\sigma = \frac{dq}{dS}$ $\frac{\text{C}}{\text{C/m}^2 \text{ m}^2}$
Electric displacement		$\mathfrak{D} = \sigma = \varepsilon \varepsilon_0 E$ $\frac{\text{C/m}^2}{\text{C/m}^2} \frac{\text{C/Vm}}{\text{V/m}}$
Electric induction	$D = 4\pi \frac{\sigma}{\sqrt{\text{dyne/cm}}}$ $= \varepsilon \frac{E}{\sqrt{\text{dyne/cm}}}$	
Electric flux	$N = \int \frac{D \cos \alpha dS}{\sqrt{\text{dyne/cm}}}$ $\frac{\text{cm}^2}{\text{cm}^2}$	$\mathfrak{N} = \int \mathfrak{D} \cos \alpha dS$ $\frac{\text{C}}{\text{C/cm}^2 \text{ m}^2}$
Gauss-Ostrogradsky theorem	$\oint \frac{D \cos \alpha dS}{\sqrt{\text{dyne/cm}}} = \sum_{\text{inside C a surface}} 4\pi q_i$	$\oint \mathfrak{D} \cos \alpha dS = \sum_{\text{inside C a surface}} q_i$ $\frac{\text{C/m}^2 \text{ m}^2}{\text{C}}$
Electric potential	$\varphi = \frac{U}{q}$ $\frac{\text{erg}}{\sqrt{\text{dyne}}}$	$\varphi = \frac{U}{q}$ $\frac{\text{J}}{\text{C}}$
Electric field intensity of point charge	$E = \frac{q}{\varepsilon r^2}$ $\frac{\sqrt{\text{dyne/cm}}}{\text{cm}^2}$	$E = \frac{q}{4\pi \varepsilon \varepsilon_0 r^2}$ $\frac{\text{C}}{\text{V/m} \frac{\text{C}}{\text{V m}} \text{ m}^2}$
Potential of the electric field of a point charge	$\varphi = \frac{q}{\varepsilon r}$ $\frac{\sqrt{\text{dyne}}}{\text{cm}}$	$\varphi = \frac{q}{4\pi \varepsilon \varepsilon_0 r}$ $\frac{\text{C}}{\text{V} \frac{\text{C}}{\text{V m}} \text{ m}}$

Continued

	Physics formulas	Elec. eng. formulas
Capacitance	$\sqrt{\text{dyne}} \text{ cm}$ $C = \frac{q}{\Phi_1 - \Phi_2}$ $\text{cm} \quad \sqrt{\text{dyne}}$	$C$ $C = \frac{q}{\Phi_1 - \Phi_2}$ $F \quad V$
Energy of the electric field of a condenser	$\text{dyne cm}^2$ $W_{el} = \frac{q^2}{2C}$ $\text{erg} \quad \text{cm}$	$C^2$ $W_{el} = \frac{q^2}{2C}$ $J \quad F$
Energy density of an electric field	$\text{dyne/cm}^2$ $w_{el} = \frac{\epsilon E^2}{8\pi}$ $\text{erg/cm}^3$	$\frac{1}{2} \epsilon \epsilon_0 E^2$ $J/m^3 \quad C/V \text{ m } V^2/m^2$
Interaction force between two charged surfaces	$\text{dyne/cm}^2$ $F = \frac{2\pi \sigma^2}{\epsilon} S$ $\text{dyne} \quad \text{cm}^2$	
Electric dipole moment	$p = ql$ $\sqrt{\text{dyne}} \text{ cm}^2 \quad \sqrt{\text{dyne}} \text{ cm cm}$	
Polarisation vector	$P = \frac{\epsilon - 1}{4\pi} E$ $\sqrt{\text{dyne}}/\text{cm} \quad \sqrt{\text{dyne}}/\text{cm}$	
Magnetic moment of a current ring	$\sqrt{\text{dyne}} \text{ cm}^2$ $M = \frac{1}{c} ISn$ $\sqrt{\text{dyne}} \text{ cm/sec cm}^2$ $\text{cm/sec}$	$M = IS n$ $A \text{ m}^2 A \text{ m}^2$
Torque acting on a current loop in a magnetic field	$N = [MB]$ $\text{dyne cm} \quad \sqrt{\text{dyne}} \text{ cm}^2 \text{ Gs}$	$N = [MB]$ $N \text{ m } A \text{ m}^2 \frac{V \text{ sec}}{m^2}$
Potential energy of a magnetic moment in a magnetic field	$U = -BM$ $\text{erg} \quad \text{Gs} \sqrt{\text{dyne}} \text{ cm}^2$	$U = -B M$ $J \quad \frac{V \text{ sec}}{m^2} A \text{ m}^2$
Ampere's law	$\sqrt{\text{dyne}} \text{ cm/sec}$ $dF = \frac{I}{c} [dl B]$ $\text{cm} \quad \text{cm Gs}$ $\text{sec}$	$dF = I [dl B]$ $N \quad A \text{ m } \frac{V \text{ sec}}{m^2}$

Continued

	Physics formulas	Elec. eng. formulas
Lorentz force	$\sqrt{\text{dyne}} \text{ cm}$ $f = \frac{e}{c} [vB]$ $\text{dyne} \frac{\text{cm}}{\text{sec}} \frac{\text{cm}}{\text{sec}} \text{Gs}$	
Magnetic flux	$\Phi = \int \frac{B \cos \alpha dS}{\text{Gs cm}^2}$ M	$\Phi = \int \frac{B \cos \alpha dS}{\frac{\text{V sec m}^2}{\text{m}^2}}$ V sec
Magnetic mass	$m = \frac{1}{4\pi} \frac{\Phi}{M}$	$m = \Phi$ V sec V sec
Magnetic field intensity created by a current element	$\frac{\text{Oe}}{\sqrt{\text{dyne}}} = \frac{I}{c} \left[ \frac{dl}{r^2} \frac{r}{\text{cm}} \right]$ $\frac{\text{cm}}{\text{sec}} \frac{\text{cm}}{\text{sec}} \text{cm}^2$	$dH = \frac{I}{4\pi r^2} \left[ \frac{dl}{r} \frac{r}{\text{m}} \right]$ A/m m <sup>2</sup>
Magnetic field of a permanent magnet	$M = m l$ M cm M cm	$M = m l$ V sec m V sec m
Relationship between $B$ and $H$	$B = \mu H$ Gs Oe	$B = \mu \mu_0 H$ $\frac{\text{V sec/m}^2 \text{ ohm sec}}{\text{m}} \text{ A/m}$ (1 ohm sec = 1H)
Interaction force between two magnetic masses (Coulomb law)	$F = \frac{m_1 m_2}{\mu r^2}$ dyne $\frac{\text{cm}}{\text{cm}^2}$	$F = \frac{m_1 m_2}{4\pi \mu \mu_0 r^2}$ N $\frac{\text{V sec V sec}}{\frac{\text{H}}{\text{m}} \text{ m}^2}$
Interaction force between two parallel currents	$dF = \frac{\mu}{c^2} \frac{I_1 L_2}{r^2} \frac{dl_1 dl_2}{\text{sec}^2}$ dyne $\frac{\text{cm}^2 \text{ cm}^2}{\text{sec}^2}$	$dF = \mu_0 \mu \frac{I_1 I_2 dl_1 dl_2}{4\pi r^2}$ N $\frac{\text{A A m m}}{\frac{\text{H}}{\text{m}} \text{ m}^2}$

Continued

	Physics formulas	Elec. eng. formulas
Magnetic potential along closed path (law of total current)	$\oint H dl = \frac{4\pi}{c} \sum I$ $\frac{\text{Oe cm}}{\frac{\text{cm}}{\text{sec}} \sqrt{\text{dyne}} \frac{\text{cm}}{\text{sec}}}$	$\oint H dl = \sum \frac{I}{A}$ $\frac{\frac{\text{A}}{\text{m}}}{\text{m}}$
Magnetic field intensity of a solenoid	$H = \frac{4\pi}{c} \frac{n}{L} I$ $\frac{\text{Oe}}{\frac{\text{cm}}{\text{sec}} \sqrt{\text{dyne}} \frac{\text{cm}}{\text{sec}}}$	$H = \frac{n}{L} \frac{I}{A}$ $\frac{\frac{\text{A}}{\text{m}}}{\text{m}}$
Emf of electromagnetic induction	$\mathcal{E}^{ind} = - \frac{1}{c} \frac{d\Phi}{dt}$ $\frac{\sqrt{\text{dyne}}}{\frac{\text{cm}}{\text{sec}} \text{ sec}}$	$\mathcal{E}^{ind} = - \frac{d\Phi}{dt}$ $\frac{\text{V sec}}{\text{sec}}$
Magnetic susceptibility	$\kappa' = \frac{\mu - 1}{4\pi}$	$\kappa = \mu - 1$
Magnetisation vector	$J = \kappa' \frac{H}{\sqrt{\text{dyne}} \frac{\text{cm}}{\text{cm}}}$	$J = \frac{\mu_0 \kappa H}{\frac{\text{V sec}}{\text{m}^2} \frac{\text{H}}{\text{m}} \frac{\text{A}}{\text{m}}}$
Maxwell's equations	$\oint E dl = - \frac{1}{c} \frac{d\Phi}{dt}$ $\frac{\sqrt{\text{dyne}}}{\frac{\text{cm}}{\text{sec}} \text{ cm}}$ $\frac{\text{M}}{\frac{\text{cm}}{\text{sec}} \text{ sec}}$ $\oint H dl = \frac{4\pi}{c} \left( I + \frac{dN}{dt} \right)$ $\frac{\sqrt{\text{dyne}} \text{ cm}}{\frac{\text{cm}}{\text{sec}} \frac{\text{cm}}{\text{sec}}}$	$\oint E dl = - \frac{d\Phi}{dt}$ $\frac{\text{V sec}}{\frac{\text{V}}{\text{m}} \text{ sec}}$ $\oint H dl = I + \frac{dN}{dt}$ $\frac{\frac{\text{A}}{\text{m}} \frac{\text{cm}}{\text{m}}}{\frac{\text{C}}{\text{A}} \frac{\text{sec}}{\text{sec}}}$
Current density	$j = \frac{I}{S}$ $\frac{\sqrt{\text{dyne}} \text{ cm/sec}}{\frac{\text{cm}}{\text{sec}} \text{ cm}^2}$	$J = \frac{I}{S}$ $\frac{\frac{\text{A}}{\text{m}^2}}{\text{m}^2}$

Continued

	Physics formulas	Elec. eng. formulas
Ohm's law	$\frac{I}{\sqrt{\text{dyne}} \frac{\text{cm}}{\text{sec}}} = \frac{\sqrt{\text{dyne}}}{R \frac{\text{sec}}{\text{cm}}}$ $\frac{j}{\sqrt{\text{dyne/sec cm}} = \lambda E}$ $\text{sec}^{-1} \sqrt{\text{dyne/cm}}$	$I = \frac{U}{R}$ <p>A ohm</p> $j = \lambda E$ $\frac{\text{A}}{\text{m}^2} \text{ ohm}^{-1} \text{m}^{-1} \text{ V/m}$
Inductance		$L = \frac{\Phi}{I}$ <p>H A</p>
Magnetic energy		$W = \frac{1}{2} L I^2$ <p>J H A<sup>2</sup></p>
Magnetic energy density	$w = \frac{\mu H^2}{8\pi}$ $\frac{\text{erg}}{\text{cm}^3}$	$w = \frac{\mu_0 \mu H^2}{2}$ $\frac{\text{J}}{\text{m}^3}$
Poynting vector	$K = \frac{c}{4\pi} \times$ $\frac{\text{erg/cm}^2 \text{ sec}}{\times [E \quad H]}$ $\frac{\sqrt{\text{dyne}}}{\text{cm}} \frac{\sqrt{\text{dyne}}}{\text{cm}}$	$K = \frac{[E \quad H]}{\frac{\text{J}}{\text{m}^2 \text{ sec}} \frac{\text{V}}{\text{m}} \frac{\text{A}}{\text{m}}}$
Mass of a unit volume of electro-magnetic field	$\frac{\text{erg}}{\text{cm}^3}$ $m = \frac{w}{c^2}$ $\frac{\text{g}}{\text{cm}^3} \frac{\text{cm}^2}{\text{sec}^2}$	

# Subject Index

- Absolute temperature, 113
- Absorption, 250-252
- Acceleration, 15
  - average, 16
  - centripetal, 17
  - Coriolis, 29
  - instantaneous, 16
  - normal, 17
  - tangential, 17
- Accelerators of charged particles, 352-353
- Additivity of molecular refraction, 524
- Amplifiers, 336
- Amplitude, wave, 257
- Antineutrino, 456
- Antinode, 97
- Antiparticles, 456-459
- Artificial radioactive products, 447-448
- Atomic nucleus, 420-449
  - mass and energy of, 427-429
- Average characteristics, 156
- Avogadro's law, 145
- Avogadro's number, 144
- Beam(s):
  - of charged particles, 340-341
  - diffracted, 295
  - intensity of, 295
- Betatron, 354
- Biological macromolecules, 493-494
- Bit(s):
  - carry, 571
  - sum, 571
- Body:
  - perfectly black, 330
  - emissive power of, 330
  - perfectly rigid, 55
- Bohr magneton, 386
- Bond(s):
  - chemical, 395-397
  - homopolar, 395
  - ionic, 395
- Brownian motion, 157
- Capacitance, 177
- Carnot cycle, 137
- Cavitation, 111
- Centres(s):
  - of gravity, 46
  - of inertia, 46
  - of mass, 46
- Cherenkov radiation, 423
- Chamber(s):
  - cloud, 420-421
  - ionisation, 421-422
  - streamer, 421
  - track, 420-421
- Charge(s):
  - interaction energy of, 185
  - point, 175
  - self-energy of, 185
- Chladni figures, 100
- Circle, tangential, 16
- Coefficient:
  - absorption, 411
  - of diffusion, 159
  - of mutual induction, 235
  - reflection, 94
  - of self-diffusion, 167
  - of self-induction, 232
  - of thermal conductivity, 160
  - of thermal output, 164
  - of viscosity, 161
- Coherence, 263-267

- Collision(s), 47
  - ideally elastic, 48
  - noncentral, 48
- Compressibility, 117
- Concentration gradient, 159
- Conduction, dependent, 356
- Conductivity, 542-544
  - n*-type, 549
  - p*-type, 549
  - thermal, 160-161
  - thermometric, 163
- Constant:
  - Boltzmann's, 144
  - Kerr, 309
  - Planck's, 248
  - radioactive-decay, 438
  - resistance, 69
  - restoring force, 65
  - Rydberg, 379
  - time, 70
- Continuous recoil, 52
- Coulomb's law, 19
- Counter(s):
  - arrangement of, 424
  - Cherenkov, 423-424
  - Geiger, 422
  - ionisation, 421
  - scintillation, 422-423
- Cross-section, effective, 142
- Crystal(s):
  - domain structure of, 535-536
  - ferroelectric, 526-528
  - liquid, 491-492
  - molecular, 471-474
  - pyroelectric, 526
  - thermal vibrations, 478-480
- Crystal symmetry, 467-470
- Curie temperature, 526
- Current, displacement, 227
- Curve(s):
  - absorption, 411
  - condensation, 496
  - crystallisation, 496
  - distribution, 147
  - fusion, 496
  - phase equilibrium, 496
  - potential, 42
  - sublimation, 496
  - transformation, 496
  - vaporisation, 496
- Cycle, Carnot, 137
- Cyclotron, 352
- de Broglie formula, 368
- de Broglie hypothesis, 368
- Defect mass, 321, 428
- Deflection of an atomic beam in a magnetic field, 384-388
- Deformation(s), 512-518
  - elastic, 512
  - homogeneous, 525
  - plastic, 513
- Degrees of freedom, 32
- Density, 87
  - electric energy, 184
  - electron, 399
  - of radial distribution of atoms, 487
- Diamagnetism, 529-531
- Dielectrics, 529-531
  - polarised, 191
- Diffraction:
  - of electrons, 367
  - of X-rays by crystals, 292-298
- Diffraction grating, 284-288
- Diffusion, 159
  - in solids, 510
- Dipole(s):
  - elementary, 243-244
  - radiation pattern of, 245
  - of a system of charges, 189
- Discharge(s):
  - arc, 358
  - in gases, 356-358
  - glow, 358
  - self-maintained, 358
  - silent, 358
  - spark, 358
- Dispersion, 250-252
  - anomalous, 251
- Displacement, chemical, 413
- Distribution:
  - Boltzmann's, 149, 541
  - statistical, 146-147
  - velocity, of molecules, 151-152
- Division of a light field into two waves, 301-302
- Doppler effect, 95-96
- Effect:
  - Doppler, 95-96
  - of Earth rotation on its form, 28
  - of Earth rotation on motion of a body on Earth's surface, 29



## Effect

- Joule-Thomson, 131
- photoelectric, 326-327, 552-553

Effective cross-section, 142

Electric susceptibility, 193

Electrodynamic action, 207

Electroluminescence, 337

- of semiconductors, 556-557

Electromagnetic action, 206

Electromagnetic field, 241

- momentum and pressure of, 241-242

Electron(s):

- free, 537
- relativistic theory of, 453-455

Electron gun, 340

Electron volt, 339

Electronic arithmetic, 569-572

Electronic computers, 568-569

Electron and ion projectors, 347-348

Electronic memory, 572-575

Electrostriction, 103, 194

Elementary particles, 450-462

- properties of, 450-462

Emission:

- of electrons, 549-551
- secondary electron, 551
- thermionic, 550

Energy, 36, 320

- activation, 436
- binding, 321
- electric, 183-185
- electromagnetic, 238-241
- free, 501
- internal, 113-114
- ionisation, 380
- kinetic, 36
- potential, of elasticity, 36-37

transformations of, 68

Energy flux, 87

Energy level(s):

- of a hydrogen atom, 379-381
- in a solid, 538-540

Enthalpy, 131

Entropy, 132

Equation(s):

- Clapeyron-Clausius, 511
- Dirac, 456
- Newton's, 30
- Maxwell's, 229
- of motion, 14
- Schrödinger's, 369
- of state, 117

Equation

- van der Waals, 120
- wave, 85, 256

Equilibrium, 42

- heat, 112

Experiment(s):

- Lebedev's 242
- Vavilov, 328
- Yoffe-Dobronravov, 327

Fabry and Perot interferometer, 274

Feedback, 75

Ferromagnetism, 532-536

Field (s):

- cylindrically radial, 179
- electric, 170-197
- energy of, 184
- force, 19
- gravitational, 19
- magnetic, 198-223
- magnetic energy of, 234-236
- on the surface of a metal object, 182
- uniform, 180
- of a uniformly charged sphere, 179

Fluctuation(s), 141, 157

- in luminous flux, 327-328

Fluorescence, 336

Flux:

- electric, 171
- energy, 87
- magnetic, 202

Force(s), 18

- acting on a moving charge, 201
- centrifugal, 25
- centripetal, 25
- coercive, 222
- electric, 186-188
- electromagnetic, 19
- external, 71
- on a freely hanging load, 24
- gravitational, 18
- Lorentz, 202
- magnetomotive, 209
- nuclear, 19
- potential, 40
- restoring, 65
- thermoelectromotive, 558-560
- "weak" interaction, 19

Forced vibrations of rods and plates, 102

- Formula(s):
  - barometric, 150
  - Boltzmann's, 151
  - Rayleigh-Jeans, 333
  - Richardson, 551
  - Stefan-Boltzmann, 331
  - Stokes', 166
  - Tsiolkovsky's, 53
  - universal, for potential, 177
  - Wiedemann-Franz, 544
- Free axes, 62
- Free vibrations, 98
  - of a rod, 98-100
    - fixed at both ends, 98
    - fixed at one end, 98
    - free at both ends, 98
  - of two-dimensional and three-dimensional systems, 100-102
- Frequency:
  - beat, 77
  - fundamental, 79
  - magnetic resonance, 412
- Function(s):
  - Fermi-Dirac, 541
  - heat, 131
  - potential, 550
  - wave, 256
  - work, 550
- Fundamental law of mechanics, 20
- Gas(es):
  - actual, 120
  - electron, 540-542
  - ideal, 119, 140
  - internal energy of, 145
  - ionised, 355
  - kinetic theory of, 140-158
  - liquefaction of, 505
  - ultra-rarefied, 168
- Gas lasers, 415-419
- Generation of wave motion, 84
- Gradient:
  - concentration, 159
  - temperature, 160
  - of velocity, 161
- Gyroscope, 64
- Half-life, 438
- Harmonics, 79
- Heat, 112-121
  - reduced, 132
- Heat equilibrium, 112
- Hologram, 290
- Holography, 289-291
- Hysteresis, magnetic, 221-223
- Induced impulse, 214
- Inductance, 232
- Induction:
  - electric, 172
  - magnetic, 199
- Inertia, moment of, 56
- Intensity of magnetic field, 205
- Interaction:
  - charge-charge, 190
  - charge-dipole, 190
  - charge-quadrupole, 191
  - dipole-dipole, 190
  - energy of, 184
  - strong, 458
  - weak, 458
- Interference:
  - in a plate, 267
  - practical applications of, 270-274
- Inversion, point of, 131
- Ionisation energy, 380
- Ionisation potential, 380
- Isospin, 433
- Joule heat
- Joule-Thomson effect, 131
- Junction:
  - metal-semiconductor, 554
  - $p$ - $n$ , 554
- Kelvin temperature scale, 113
- Kerr constant, 309
- Kinetic energy, 36
- Kinetic theory of gases, 140-158
- Kirchhoff's law, 329
- Lande factor, 531
- Laser, 336
  - absolute maximum efficiency of, 417

## Law(s):

- Ampère's, 201
- Avogadro's, 145
- Boltzmann's, 148-149, 409, 541
- Boyle-Mariotte, 128
- of conservation of mechanical energy, 40-41
- of conservation of momentum, 45
- of constant angles, 464
- Coulomb's, 19
- of electromagnetic induction, 212-213
- Faraday's, 213
- fundamental, of mechanics, 20
  - application of, to accelerated rectilinear motion, 22-24
  - application of, to circular motion, 25-27
- general, of chemical and nuclear transformations, 435-437
- generalised, of induction, 225
- Kirchhoff's, 329
- Lenz's, 226
- of mechanics, in a noninertial system of coordinates, 21
- Mendeleev periodic, 389
- Newton's, 20
- Ohm's, 231, 543
  - differential form of, 231
- statistical, 147
- Stefan-Boltzmann, 331
- Layer, barrier, 554-555
- Lens, electron, 341-343
- Light:
  - natural, 254
  - polarised, 254
  - propagation of, in uniaxial crystals, 301-304
- Line(s):
  - of force, 171
  - Raman, 409
- Lissajous figure, 81
- Logarithmic decrement, 69
- Luminescence, 336-337

- Magnetic hysteresis, 221-223
- Magnetic permeability of vacuum, 205
- Magnetic substance, 529-536
- Magnetohydrodynamics, 360
- Magnetogas dynamics, 360
- Mass, 319

- Mass defect, 321, 428
- Mass spectrograph, 351

## Material(s):

- piezoelectric, 525
- pyroelectric, 525
- seignette-electric, 193
- Mean free path, 141-142
- Mechanics, quantum, 369
- Mendeleev periodic law, 389
- Mesons, 452-453
- Meson theory of nuclear interaction, 451-452

## Method(s):

- of analysing observations, 425-426
- Chladni, 100
- Debye, 297
- experimental, of nuclear physics, 420-426
- Laue, 297
- nuclear emulsion, 424-425
- parallelogram, 14
- powder, 297
- of X-ray analysis, 296-298
- Microelectronic circuits, 562-564
- Microparticles (s):
  - wave properties of, 367-378
- Microprocessors, 566-568
- Microscope:
  - electron, 343-347
  - resolving power of, 344
- Molecule (s), 395-419
  - diatomic, 405
  - electron cloud of, 399-401
  - energy levels of, 401-402
  - geometries of, 397-399
  - rotational spectrum of, 402-404

## Moment:

- electric, of dipole, 189
- of force, 58
- of inertia, 56
- magnetic, 198, 388
- of momentum, 60

## Momentum, 45

- angular, 60
- conservation of, 45

## Monocrystals, 463-464

## Motion:

- horizontal, under the action of a constant force, 22
- relative, 51
- relativity of, 20

- Neutrino emitted in beta-decay, 434-435  
 Nicol prisms, 305  
 Node, 97, 464  
 Nonelastic impact, 47  
 Nucleus:  
   atomic, 420-449  
   radius of, 432  
   spin and magnetic moment of, 430  
 Number:  
   Aliven, 361  
   Avogadro's, 144  
   Hartmann, 362  
   Loschmidt's, 145  
   Mach, 360  
   quantum, 374  
   Reynolds, 165  
   spin quantum, 387  
   Stuart, 362  
   wave, 403
- Octupole, 189  
 Optical activity, 310-311  
   basic theory of, 311-313  
 Oscillation(s):  
   coherent, 263  
   electric, 236-238  
   self-sustained, 75  
 Overtone, 79
- Paramagnetism, 531-532  
 Particle(s), 456-459  
   in electric field, 338  
   elementary, 450  
   in magnetic field, 339  
   nuclear, 426  
   parity of, 433  
 Peltier effect, 560  
 Pendulum, 67  
   mathematical, 67  
   period of, 67  
   physical, 68  
 Period:  
   characteristic, 66  
   half-life, 438  
   natural, 66  
   of vibration, 65  
 Permeability:  
   relative: 205
- Permeability  
   magnetic, in vacuum, 205  
 Permittivity, 172  
 Phase diagrams, 495-497  
 Phase diagram and properties of helium,  
   497-500  
 Phase stability 353-354  
 Phase transformations, 495-511  
 Phosphor, 423  
 Phosphorescence, 336  
 Photoeffect, 552  
   extrinsic, 552-553  
   intrinsic, 553  
 Photoluminescence, 337  
 Photon, 324, 457  
 Piezoelectric effect, 103  
   inverse, 103  
 Plasma, 356, 359  
   in a magnetic field, 360  
 Point of inversion, 131  
 Poise, 161  
 Polarisability, anisotropic, 299-301  
 Polarisation, 192, 254  
   of crystal substances, 193  
   elliptical, 302-303  
   molecular, 520  
   of polar and nonpolar molecules, 521-523  
 Polarisers, 305-306  
 Polaroids, 306  
 Polycrystalline materials, 515  
   mechanical properties of, 515  
 Polycrystalline substances, 463-464  
 Polymers, 492-493  
 Postulate:  
   Clausius', 139  
   Thomson's, 139  
 Potential(s):  
   constant, 555-556  
   electric field, 174  
   ionisation, 380  
   thermodynamic, 501  
 Potential energy, 36  
   elastic, 37  
   of electrical interaction of charges, 39  
   gravitational, 38  
 Precession, 64  
 Principle:  
   of constancy of the velocity of light, 315-  
     317  
   of entropy existence, 132  
   of equivalence, 321-323  
   Huygens', 92

## Principle

- Huygens-Fresnel, 92
- of increasing entropy, 134
- Michelson interferometer, 273
- of operation of a heat engine, 136
- Pauli exclusion, 383-384
- uncertainty, 370

## Probability of a state, 153

## Processes:

- adiabatic, 113, 127
- cyclic, 124
- diffusion, 159
- isobaric, 125
- isochoric, 124
- isothermal, 126
- Joule-Thomson, 130
- steady, 163
- thermodynamic, 122-131

## Pumping, 415

## Quadrupole, 189

## Quantum-mechanical oscillator, 336

## Quantum mechanics, 369

- fundamental concepts of, 368-370
- fundamental law of, 369

## Quantum nature of a field, 324-337

## Quantum number, 381-382

## Quarks, 462

## Radar, 260-261

## Radiation:

- black-body, 330-332
- Cherenkov, 423
- electromagnetic, 243-249
- stimulated emission of, 335-336
- thermal, 332-335

## Radiator(s):

- primary, 243
- secondary, 243

## Radioactivity, 437-441

- artificial, 438
- natural, 438

## Radius:

- of curvature, 16
- electron, 185

## Random events, 146

## Reaction(s):

- chain, 443-445

- endothermic, 437
- exothermic, 437
- fission, 441-443
- nuclear, 440-441
- photonuclear, 440
- thermonuclear, 448-449

## Reactor(s):

- breeder, 447
- nuclear, 445

## Recoil, 52

## Reflection of waves, 92

## Refraction:

- double, 299-313
- index of, 250
- molecular, 520
- of waves, 92

## Refractive index, 93

## Relativity, general theory of, 314

## Relativity principle, Galileo's, 21

## Resistance:

- acoustic, 87
- inertial, 22
- wave, 87

## Resonance, 71

- magnetic, 412-413
- quadrupole, 413-415

## Rogovsky belt, 214

## Scattering, 275-291

- in a nonhomogeneous medium, 282-284

## Self-diffusion, 167

## Semiconductors, 546-547

- properties of, 546

## Series, radioactive, 438

## Slippage, 513

## Sound:

- intensity and loudness of, 106-107
- objective and subjective nature of, 105

## Space lattice, 464-470

## Spectrum(a)

- absorption, 391, 410-411
- of atomic nuclei, 432-434
- baryon, 459
- electromagnetic, 247
- emission, 391
- Raman, 408
- vibration, 78
- X-ray, 392-394

## Spectrum analysis, 79

## Sphere, Fermi, 540

- Spin, 387
- Spiral dislocation, 508
- Square well, 373
- State:
  - degenerate, 377
  - metastable, 502
  - probability of, 153
- Statistical law, 147
- Statistics:
  - Boltzmann, 542
  - Bose-Einstein, 334, 541
  - Fermi-Dirac, 541
- Superconductivity, 544-546
- Susceptibility:
  - electric, 193
  - magnetic, 217
- Synchrocyclotron, 353
- Synchrotron:
  - electron, 355
  - proton, 354
- Systems:
  - closed, 41
  - microscopic, 115
- System of units:
  - CGS, 32
  - FLT, 33
  - SI, 33
- Temperature, 112-121
  - Curie, 526
- Theorem:
  - Gauss, 179
  - Gauss-Ostrogradsky, 177
  - Nernst's, 133
  - Poynting's, 239
- Theory of relativity, 314-323
- Thermal capacity:
  - at constant pressure, 125
  - at constant volume, 125
- Thermodynamics:
  - first law of, 115
  - second law of, 135, 139
  - third law of,
- Thermometric coefficient,
  - of change of pressure, 117
  - of dilation, 117
- Thomson effect, 560
- Thomson heat, 561
- Torque, 58
- Total mechanical energy, 41
- Transformation:
  - Martin, 509
  - phase, 495-511
- Transistors, bipolar, 562
- Tube, electron-beam, 348-350
- Tunnelling through a barrier, 377-378
- Ultimate strength, 515
- Ultra-rarefied gases, 168
- Ultrasonics, 111
- Umov-Poynting vector, 239
- Uncertainty principle, 370
- Uniform field, 180
- Universal formula for potential, 177
- Vector(s):
  - of magnetic induction, 199
  - magnetisation, 216
  - polarisation, 192
  - Poynting, 239
- Velocity:
  - angular, 55
  - average, 14
  - instantaneous, 15
  - linear, 55
  - root-mean-square, 144
- Vibration(s):
  - damped, 68
  - forced, 71-74
  - free, 98
  - mutually perpendicular, 80
  - parallel, 76
  - piezoelectric, 103-104
  - of polyatomic molecule, 406
  - pressure and velocity of, 86
  - self-sustained, 74
- Viscosity, 161
  - kinematic, 163
- Wave(s):
  - acoustic, 481
  - directed radiators of, 288-289
  - elastic, 89
  - electromagnetic, 250
    - propagation of, 250-261
  - interference of, 90-92

- Wave(s)  
  path-difference of, 91  
  radio,  
    propagation of, 258  
  reflected, 92  
  refracted, 92  
  spherical, 92  
  standing, 97-104  
  thermal, 480-482  
  travelling, 82-96  
  ultrasonic, 90  
Wave equation, 85  
Wave front, 88  
Waveguide, 403  
Wavelength, 86  
Weight of a body, 28  
  
Work, 34  
  rotational, 58  
  
X-ray analysis, 296-298  
  
Yoffe-Dobronravov experiment, 327  
  
Zero, absolute, 113  
Zone, 464  
  axis of, 464

TO THE READER

Mir Publishers would be grateful for your comments on the content, translation and design of this book. We would also be pleased to receive any other suggestions you may wish to make.

Our address is:

Mir Publishers

2 Pervy Rizhsky Pereulok

I-110, GSP, Moscow, 129820

USSR



# Physics. A General Course, Vols. I-II

I. SAVELYEV, D. Sc.

This modernized course in physics sums up the experience gained by the author during many years of teaching the subject at the Moscow Physical Engineering Institute. The accent is placed not on imparting information, but on the formation of physical thinking by students and on their mastering the ideas and methods of the science of physics. Methodically more improved ways of treating a number of questions have been found. This has made the expounding of the material stricter, and at the same time simpler and easier to understand.

The book is intended first of all for students of higher educational institutions with a broad syllabus in physics.

The material is arranged, however, so that by omitting separate portions, it can be used by students taking an ordinary course in physics.

The third volume of the book, covering quantum optics, atomic physics, physics of solids, and physics of the atomic nucleus and elementary particles, will be published in 1981.

# Thermodynamics, Statistical Physics and Kinetics

Yu. RUMER, D. Sc., M. RYVKIN, Cand. Sc.

This book is meant for readers embarking on a theoretical course in thermodynamics, statistical physics and kinetics. It is thus assumed that the reader possesses a sufficient background in general physics, higher mathematics and quantum mechanics. The main aim of this book is to acquaint the reader with methods of thermodynamics, statistical physics and kinetics starting from elementary concepts and bring him up to a level where he can follow monographs and articles, and independently tackle problems. With this end in view, elementary questions have been treated with the same importance as quite advanced and fairly complicated topics. Such complicated parts have been marked with asterisks and may be omitted on first reading.

The book contains a large number of problems and exercises, and is intended for students of physics and mathematics at the universities.

**CONTENTS.** **Thermodynamics.** General Laws of Thermodynamics. Systems with a Variable Quantity of Matter. Phase Transitions. **Statistical Physics.** Statistical Distribution for Ideal Gases. The Maxwell-Boltzmann Gas. Degenerate Gases. Systems of Interacting Particles. The Gibbs Method. Theory of Fluctuations. Phase Transitions. **Elements of Kinetics and Non-Equilibrium Thermodynamics.** Kinetics. Elements of Non-Equilibrium Thermodynamics. Mathematical Appendix.







